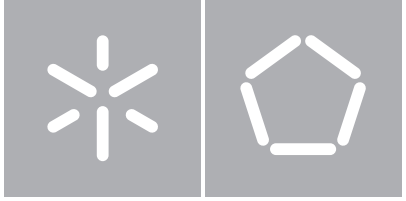
Universidade do Minho
Escola de Engenharia

Adrien Fernandes Machado

Finding new genes and pathways involved in cancer development by analysing insertional mutagenesis data

Janeiro de 2016



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Adrien Fernandes Machado

Finding new genes and pathways involved in cancer development by analysing insertional mutagenesis data

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho realizado sob orientação de

Dr. Jeroen de Ridder
Dr. Isabel Rocha

Janeiro de 2016

Anexo 3

Declaração

Nome Adrien Fernandes Machado

Endereço Eletrónico adrienfmachado@gmail.com

Número do Cartão de Cidadão 13909954

Título da Dissertação Finding new genes and pathways involved in cancer development by analysing insertional mutagenesis data

Orientador Professor Dr. Jeroen de Ridder

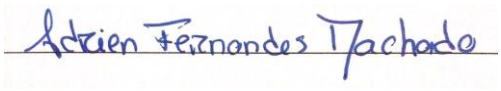
Co-orientador Doutor Isabel Rocha

Ano de Conclusão 2016

Designação do Mestrado Bioinformática

Universidade do Minho, 29 / 01 /2016

Assinatura:



"The purpose of life is to live it,
to taste experience to the utmost,
to reach out eagerly and without fear
for newer and richer experience.".

Eleanor Roosevelt (1884 - 1962)

Acknowledgement / Agradecimentos

First and foremost, I would like to thank my supervisor Dr. ir. Jeroen de Ridder for the opportunity to work this topic and for the continuous support, teaching and patience that he provided me during the this work.

A special thanks to the Delft Bioinformatics Lab, were all the work was done, for all the group meetings that enriched my knowledge about bioinformatics and great moments.

Gostaria de agradecer à Professora Isabel Rocha pelo apoio que me deu para a realização do Erasmus assim como a ajuda ao estabelecer esta nova parceria.

Toda esta jornada não seria possivel sem o apoio da minha família, dando a possibilidade de realizar os meus objetivos e permitirem-me crescer profissionalmente. Um grande Obrigado aos meus pais e à minha irmã.

For the amazing housemates Cornel, Frank, Friso and Tom! Guys, you were incredible! Thank you for your fellowship along these months!

Aqueles que me acompanharam diariamente nesta aventura na Holanda - o *gangtuga* - um muito obrigado à Sara, Fred, João, Manel, Mariana, Marina, Sofia, Fitas.

Um agradecimento aos meus membros constituintes da equipa de camaradagem do Mestrado de Bioinformática 2013/14, em especial ao Daniel, Santa, Manel, Marisa, Lima, Tania e Vitor pelos conselhos e companheirismo destes últimos 2 anos.

À Joana e à Preta, um obrigado pela camaradagem e apoio ao longo destes últimos meses.

And, last but not least, thank you babe, for everything, all the support. You know.

ABSTRACT

Cancer emerges from an uncontrollable division of the organism's cells, creating a tumour. These tumours can emerge from any part of the human body. The increase of cellular division and growth can be created by mutations in the genome. Several methodologies are approached, in the research, to finding new cancer genes. The insertional mutagenesis (IM) has been one of the most used, in which the mouse is infected by a retrovirus or a transposon, increasing the gene expression in the insertions' vicinity.

The data used in work essay are a collection of independent studies of IM in mice. After its processing, the data has 3,414 samples, having information of 7,751 genes. Each sample matches a type of cancer (colorectal, hematopoietic, hepatocellular carcinoma, lymphoma, malignant peripheral nerve sheath, medulloblastoma and pancreatic).

The main goal of this project is to determine if there are specific genes for a particular type of cancer. And, if there are, which are the 15 most evolved genes for that type of cancer.

Machine learning (ML) is a subject where its goal is to increase knowledge based on given experimental data, allowing it to execute predictions and accurate decisions. To answer our purpose, it is necessary the transform the data into a dissimilarity relation between samples. Different approaches were used: two of them are known from the literature (Hamming distance and Jaccard distance) and two new metrics were developed (Gene Dependent Method (GDM) and Gene Independent Method (GIM)). With these transformations, unsupervised learning methods (such as Principal Component Analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE)) and supervised learning approach, testing different classifiers by crossed validation, were used.

The main results show that some genes may be specific to a particular type of cancer. Therefore, it is possible to create a ranking gene, according to its importance to a type of cancer. 105 genes are presented (15 genes of each type of cancer), of which 18 were not annotated yet and 19 have already been mentioned in the literature to be involved in the development of the selected cancer tissue. Afterwards it must be performed a proper *in vitro* and *in vivo* validation.

Keywords: Cancer; cancer genes; insertional mutagenesis; machine learning.

RESUMO

O cancro surge da divisão incontrolável de células de um organismo, criando um tumor. Estes tumores podem surgir em qualquer parte do corpo do ser vivo. O aumento da divisão e crescimento celular pode dever-se a mutações no genoma. São várias as metodologias abordadas na investigação para a descoberta de novos genes de cancro. A mutação por inserção (IM) tem sido uma abordagem bastante utilizada, no qual o rato é infetado por um retrovírus ou um transposão, aumentando a expressão do gene que se encontra na vizinhança da inserção.

Os dados usados neste trabalho correspondem a uma coleção de estudos independentes de IM em ratos. Após o seu processamento, os dados contêm 3,414 amostras, tendo informação de 7,751 genes. Cada uma das amostras corresponde a um tipo de cancro (colorectal, tecido hematopoiético, carcinoma hepatocelular, linfoma, tumor maligno de bainha nervosa, meduloblastoma e pâncreas).

O objetivo principal deste projeto é determinar se existem genes específicos para um determinado tipo de cancro e, se sim, quais são os 15 genes mais envolvidos para o desenvolvimento do mesmo.

A aprendizagem de máquina (ML) tem como objetivo ganhar conhecimento com base em dados experimentais fornecidos, permitindo que este possa realizar previsões e decisões precisas. Para se responder ao objetivo, é necessária a transformação dos dados numa relação de dissimilaridade entre amostras. Foram usadas quatro abordagens: duas delas são descritas na literatura (a distância de Hamming e a distância de Jaccard) e duas novas métricas foram desenvolvidas (o método de gene dependente (GDM) e o método de gene independente (GIM)). A partir destas transformações foram usadas metodologias de aprendizagem não supervisionada (a Análise de Componentes Principais (PCA) e o *t-distributed stochastic neighbor embedding* (t-SNE)), e a metodologia supervisionada, testando diferentes classificadores por validação cruzada.

Os resultados principais mostram que existem genes que poderão ser específicos para um dado tipo de cancro. Assim sendo, é possível criar uma ordenação dos genes de acordo com a sua importância face a um tipo de cancro. São apresentados 105 genes (15 genes para cada tipo de cancro), dos quais 18 ainda não foram anotados e 19 já foram mencionados na literatura por estarem envolvidos no desenvolvimento do cancro do tecido selecionado. Posteriormente deverá ser realizada a devida validação *in vitro* e *in vivo*.

Palavras-chave: Aprendizagem de máquina; cancro; genes de cancro; mutagénese por inserção.

CONTENTS

List of Figures

List of Tables

List of Acronyms

2D	two dimensions
3D	three dimensions
AUC	area under the curve
BCC	Basal cell carcinoma
bp	base pair
cDNA	complementary DNA
CIS	common insertion site
CV	cross-validation
DNA	deoxyribonucleic acid
FN	False Negative
FP	False Positive
GBM	Glioblastoma multiforme
GDM	gene dependent method
GIM	gene independent method
HCC	Hepatocellular carcinoma
HSC	Hematopoietic stem cells
ID3	Iterative Dichotomiser 3
IM	insertional mutagenesis
kb	kilobase
kNN	k-Nearest Neighbour
MATLAB	Matrix laboratory
ML	Machine Learning
MuLV	murine leukemia virus
MMTV	mouse mammary tumour virus
MNIST	Mixed National Institute of Standards and Technology
MPNST	Malignant peripheral nerve sheath tumour
NMC	Nearest Mean Classifier
PCA	Principal Component Analysis
PCR	polymerase chain reaction
RNA	ribonucleic acid
ROC	receiver operating characteristic
SCC	Squamos cell carcinoma
SVM	Support Vector Machine
T-ALL	T-cell acute lymphoblastic leukaemia
TN	True Negative
TP	True Positive
t-SNE	t-distributed stochastic neighbor embedding

MATHEMATICAL NOTATIONS

Functions

acc	Accuracy function
C	A classifier function, where $C(\widehat{x}) = \widehat{y}$
$d(a,b)$	Distance of a from b
d_E	Euclidian distance
d_{GIM}	Gene independent method distance
d_{GDM}	Gene dependent method distance
d_H	Hamming distance
d_J	Jaccard distance
ERR	Error function
F	Example of a mapping function
H	Entropy function
$I(a,b)$	Indicator function where $I(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases}$
IG	Information gain function
δ	Indicator function where $\delta(a,b) = \begin{cases} 1 & \text{if } a \neq b \\ 0 & \text{if } a = b \end{cases}$
$sgn(a)$	Signum function where $sgn(a) = \begin{cases} -1 & \text{if } a < 0 \\ 0 & \text{if } a = 0 \\ 1 & \text{if } a > 0 \end{cases}$
SSE	sum of squared errors function

Mathematical operations

$\overline{a}$	mean value of a
$\arg\max_a f(a))$	the value of a that leads to the maximum of $f(a)$
$\arg\min_a(f(a))$	the value of a that leads to the minimum of $f(a)$
$\log_2(a)$	logarithm with base 2 of a
$\sum_{i=1}^{n} a_i$	The sum function from $i = 1$ to $n : a_1 + a_2 + ... + a_n$
$\prod_{i=1}^{n} a_i$	The product function from $i = 1$ to $n : a_1 \times a_2 \times ... \times a_n$

Machine learning

$D(X, Y)$	Dataset D, with a set of vectors X and their respective label Y
D_t	Training set
X	Set of vectors: $X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_m\}$
$\vec{x}_i$	the i^{th} vector, $\vec{x}_i \in X$
$\hat{x}$	new vector to be predicted
Y	set of labels: $Y = \{y_1, y_2, ..., y_z\}$
y_i	label of the i^{th} vector, $y_i \in Y$
$\hat{y}$	the label predicted by a classifier C, $C(\hat{x}) = \hat{y}$
X^{y_1}	Set of vectors belonging to the class y_1: $X^{y_1} = \{\vec{x}_1^{y_1}, \vec{x}_2^{y_1}, ..., \vec{x}_m^{y_1}\}$
$\vec{x}_i^{y_1}$	the i^{th} vector of the class y_1, $\vec{x}_i^{y_1} \in X^{y_1}$
A	Set of attributes (or features), $A = \{a_1, a_2, ..., a_z\}$
a_i	the i^{th} attribute, $a_i \in A$
A'	Set of mapped attributes, $A' = \{a'_1, a'_2, ..., a'_{nr}\}$, $nr < n$
m	number of vectors
m^{y_1}	number of vectors belonging to class y_1
n	number of attributes
nr	number of attributes after mapping
z	number of different labels

Probability

$P(a)$	Probability of event a
$P(a\|b)$	Probability of event a given event b

Sets

$y \in Y$	y is an element of the set Y
$A \cap B$	Intersection of the sets A and B
$A \cup B$	Union of the sets A and B
$\mathbb{R}$	Set of real numbers

Symbols

$a \equiv b$	a is equivalent to b
$\#(\vec{a} = 1)$	number of elements of $\vec{a}$ is equal to 1.

1

INTRODUCTION

1.1. MOTIVATION

Cancer is the name given to an assembly of more than 100 diseases[1]. All these diseases can be very distinctive of each other. Nonetheless, they all have a similar starting point: an abnormal cell division, creating more cells than the body needs, producing a tumour.

Every year the number of new cases of cancer increases. In 2008, 12.7 million new registrations and 7.6 million deaths as a possible result of this disease were estimated [1]. A recent study, analysing data from 2012, estimates a registration of 14.1 million new cases of cancer and 8.2 million deaths as a possible result of this disease [2]. This growth is caused essentially as a result of the populations' rise, as well as to the exposition to risk factors.

Cancer is caused by changing the genetic information - the deoxyribonucleic acid (DNA)- of a cell. This alteration is called mutation and most of the time cells can repair it.

Cancer research is extremely important due the impact it can cause on our society. Analysing the changes of a gene or pathways, it is possible to predict which patients are likely to have a better or worse diagnosis.

1.2. OBJECTIVES

The key focus of this project is to improve understanding of biological processes that lead to cancer. The data collected contains information of exogenous DNA which integrates the mouse's genome - *insertional mutagenesis (IM)*. This integration will activate genes in its vicinity, in special, cancer genes.

[1] http://www.cancer.gov/about-cancer/what-is-cancer, accessed: July 2015

The main biological question of the project is to determine which genes are likely to be a candidate as cancer genes to a specific type of cancer. To answer this question, the strategy is to use Machine Learning (ML).

Machine learning uses algorithms that can learn from data [3]. Classification methods allow to make predictions and decisions. For example, classification techniques have been used to extract cancer genes from large gene expression datasets [4–6]. IM screening data are represented by a very sparse Boolean matrix, and as such is very different from gene expression data. For this reason, the first problem is to know which classifier is suitable for application to sparse Boolean data. Several classifiers will be tested. To capture this in the classifier, the data will be transformed in a distance matrix. This evaluation will be performed using two classes, representing two distinct cancer types. To conclude, it will be evaluated feature selection methods to determine which genes interact in specific types of cancer.

1.3. STRUCTURE OF THE DISSERTATION

This dissertation is divided into seven chapters. In this first chapter, a brief introduction of the motivation and the main aims of the work are provided.

Second chapter - Cancer research

Introduces several aspects related to cancer, as well as the research done to find new cancer genes.

Third chapter - Machine Learning

Presents an explanation of several important aspects of learning algorithms and their evaluation. To understand differences in the data it is explained some approaches to transform it. It is also discusses an approach to find important features from a dataset.

Fourth chapter - Data

Describes how the data was generated, how it was organized and explains the pre-processing performed, to have the final dataset.

Fifth chapter - Methodology

Explains the several steps of the work developed: the approaches used to transform the data; the unsupervised and supervised learning methods; as well as the ranking method.

Sixth chapter - Results and discussion

Addresses the main results of this work: the visualization of the unsupervised learning methods; the performances of selected classifiers used in supervised learning methods; and a list of potential genes that are involved in tumourigenesis.

Seventh chapter - Conclusion

Describes the main conclusions of the work, the limitations and recommendations for future work.

2

CANCER RESEARCH

2.1. CANCER

Cancer is the name given to a group of diseases. All the different types of cancer arise with an unexpected aberrant cell division - *neoplasia* -, which disseminate to near tissues - *metastesis*. The Human body contains approximately 37 trillion and 200 million (3.72×10^{13}) cells and all of them can originate a tumour [7]. Not all tumours lead to cancer. In fact, tumours can be distinguished in two groups: *benign* and *malignant*. The first one does not have the ability to invade other tissues, which makes the removal a simple process. The second can spread to neighbour tissues. Even if the tumour is cut out, the organism still carries some cancer cells, which later can develop a new tumour.

Cancer is a genetic disorder. It is caused by changing gene expression, which controls the cell function. These changes can generate mutations. The probability of having a sporadic mutation in each base pair (bp) is estimated to be 1 in 100 million (1.1×10^{-8})[8]. This value may seem low, but due the enormous quantity of bp that the Human genome contains, as well as, the massive number of cells each individual has in their lifetime and their risk behaviours, the probability increases largely. In addition, there are many agents which change DNA. They can be caused naturally by environmental factors due to physical (e.g. radiation), chemical (e.g. smoke) and biological (e.g. virus) causes, as well as, by genetic alterations (sporadic or hereditary) [9, 10].

Mutations happen all the time in our cells. In fact, during the cell cycle, cells have mechanisms which can detect an error and repair them. If the cell cannot replace its damages, it will receive a signal to initiate the process to its death -*apoptosis* [11].

Not all cancer cells are generated by mutations. Epigenetics is the study of cellular and physiological alteration caused by exogenous factors. In this situation, the alterations do not change the nucleotide sequence. Epigenetics alterations can change the expression of a gene, increasing or decreasing it.

2.1.1. THE CANCER CELL EVOLUTION

The cell - the basic structural, functional and biological unit of organisms – preserves, inside it, one of the most important discovery in biological science, the DNA. This molecule contains the information that the cell needs. This information is stored in genes. One of the functions of the cell is to reproduce itself, dividing itself in two daughter cells and transmit its genetic information - *cell cycle*. It is estimated that this mechanism repeats between 50-70 billion cells per day in our organism to replace dead cells [12]. This process has two steps: *interphase* - the cell growth, accumulating compounds and duplicating its DNA; and *mitosis* - the cell splits itself into two distinct cells. These two phases have checkpoints, which ensures the appropriate replication of the DNA and division of the cell [13].

Before the transition from a normal to a cancer cell - *tumourigenesis* - can happen, the cell must overcome all its protections. It can be considered as an accelerated version of Darwin's evolution theory: the individual receives an inherited genetic variation, it gets selective advantages and transmits it to its next generation [14–16]. In fact, if genes have different activities than usual, it will change, therefore, the cell's activity and induce the accumulation of several alterations in its DNA along its generations (during years or decades). They will overcome all checkpoints and gain some selective advantages compared to the normal cells [17]. With this accumulation, the cell will change its properties, and then, can evolve to a cancer phenotype [18].

To be considered as a cancer cell, the cell has to have several characteristics. Hanahan and Weinberg [19, 20] suggest that cancer cells can be summarised in 10 hallmarks (Figure 2.1):

Six basic hallmarks, representing the fundamental basis of malignancy:

Sustaining proliferative signalling

Normal cells regulate carefully the process of growth signals, insuring the cell homeostasis. However, cancer cells can overcome this mechanism, for example, producing more growth factor, increasing the number of receptors on the cell surface and changing the signalling pathways.

Evading growth suppressors

Normal cells rely on anti-growth signals to regulate their growth. Most of these pro-

cesses depend on the actions of tumour suppressor genes. Cancer cells become insensitive to mechanisms that regulate negatively the cell proliferation.

Resisting cell death

Due the DNA damage and other cellular stresses, normal cells may initiate apoptosis [11]. Most of cancer cells are less sensitive to similar stresses, avoiding apoptosis and contributing to the uncontrollable division.

Enabling replicative immortality

The number of division a cell can do is limited. These limits are usually established by telomeres (the ends of chromosomes). Along each cell division, in normal cells, telomeres get shorter until they are not able to divide. In contrast, in cancer cells, telomeres are preserved, allowing the cell to divide an unlimited number of times.

Inducing angiogenesis

Angiogenesis is the process of creating new blood vessels, mediated mainly through vascular endothelial growth factor. It plays a critical role in tumour growth, supplying the cancer cells with oxygen and nutrients.

Activating invasion and metastasis

Metastasis is the cause of 90% of deaths from solid tumours [21]. Here, cancer cells may escape from the primary site and disseminate into distant organs. This process is not well understood, but it is known to involve a large number of secreted factors which breaks the tissue, allowing the invasion into blood vessels, and then, creating a new tumour in another place in the organism.

The acquisition of these hallmarks of cancer is made possible by two enabling characteristics:

Genome instability and mutation

Cancer cells achieve genome instability by increasing their mutation's rate. They increase their sensitivity to mutagenic agents or breakdown of DNA repair machanisms.

Tumour promoting inflammation

Immune cells might infiltrate tumours and produce inflammatory responses. Inflammation can release chemicals into the tumour microenvironment, leading to genetic mutations and helping tumours acquire the hallmarks of cancer

Furthermore, two emerging hallmarks might be involved in the development of cancer:

Figure 2.1: Hallmarks of cancer.
Characteristics the normal cell has to collect to achieve the cancerous phenotype.
Figure adapted from Hanahan and Weinberg (2011) [20].

Deregulating cellular energetics

The uncontrolled growth and division of cancer cells may rely on the reprogramming of cellular metabolism, including increased aerobic glycolysis (known as the Warburg effect).

Avoiding immune destruction

The immune system is responsible for protecting the organism, including recognition and elimination of cancer cells. Evasion of this immune surveillance by weakly immunogenic cancer cells is an important emerging hallmark of cancer.

2.1.2. THE CANCER GENES

It is widely accepted that tumourigenesis is a process which arises as a result of different activity of the genes present in the cell and they can differ between different types of cancers. The main challenge that researchers face is understanding which genes must be active or inactive to stop the normal operation of the cell and arouses to cancer. Some of their names are known [22]. However, it is believed that most of them are still a mystery. The term "cancer gene" will be used throughout this dissertation to describe a gene for which mutations have been causally implicated in cancer. Cancer genes are commonly classified in two groups:

Proto-oncogenes

Proto-oncogenes (e.g. *myc* and *ras*) are genes that incentives the cell growth. They turn to oncogene when they are mutated, being more active, allowing cells to grow more and surviving when they should not. Usually the overexpression of these genes is caused by gene amplification or chromosomal translocation [23].

Tumour suppressor genes

Tumour suppressor genes (e.g. *p53*) have as main purpose the reduction of cell proliferation. When these genes do not work correctly, the cell is able to grow out of control. This happens due to the mutation, causing loss of function of the gene.

It is important to understand that tumourigenesis develops as a result of activation of proto-oncogene, becoming an oncogene, and the inactivation of tumour suppressor genes. In general, to the tumour suppressor gene loss its function, it must be mutated in both alleles (recessive mutation)[24]. In contrast, since the mutation in oncogenes corresponds to the gain-of-function, most of its mutations involve only an individual allele (dominant mutations)[25].

2.2. THE RESEARCH

Cancer formation results from gene mutations, which regulates the cell's growth. Major tumours result either gain or loss-of-function of gene's activity. Discovering which genes are involved in tumourigenesis allows, for example, the creation of drugs that can act against this abnormal gene or the protein encoded.

2.2.1. DISCOVERING GENES AND PATHWAYS INVOLVED IN CANCER

In order to find new genes which leads to cancer's hallmarks, several strategies are used. Most of them are tested in humans and in mouse [26]. Some techniques use tumour tissues from patients. On the other hand, a large part of the research uses animal models. The mouse is the biological model most used in research. It has a fast reproduction rate, a short life cycle and a small size, so it can be preserved in smaller spaces. In addition, the mouse is also physically and genetically similar to humans. Most genetics finding in mouse have a homology in human[27].

From all methods to discover new candidates to cancer genes, insertional mutagenesis (IM) has been a very efficient tool. The following work uses this approach in mouse and it is described below.

2.2.2. INSERTIONAL MUTAGENESIS

Insertional mutagenesis (IM) is a mechanism by which an exogenous DNA element integrates the genome of a host cell. It can be used in several fields of molecular biology, such as, gene therapy [28], gene regulation[29] and oncogene discovery [26]. As mentioned before, the mouse is the most used model for cancer study, although IM has also been performed several different organisms, such as other vertebrates (e.g. chicken [30], zebrafish [31]), insects (e.g. *Drosophila melanogaster* [32]), plants (e.g. *Arabidopsis thaliana* [29] and rice [33]) and fungus [34].

HOW IT WORKS

In this technique, the mouse is infected by a retrovirus or a transposon. They will infect the healthy cells and integrates their genome in the host cells. By consequence, this integration will deregulate genes in the vicinity, even in large distances [35], and can cause a perturbation of the phenotype. When the incorporation increases the expression of proto-oncogenes or decreases of tumour suppressor genes, it can result in an accelerated cell proliferation. The integration can alter the gene expression in different ways: either up- or downstream, changing its expression level and rarely the encoded protein; or within the gene, resulting in a different encoded protein or in its inactivation [36]. Regions in the genome that contain constantly insertions located in the same *loci*, in independent tumours, are referred as common insertion site (CIS). CIS show a significant overlap with human cancer genes (Figure 2.2)[37].

MECHANISM OF INSERTIONAL MUTAGENESIS

In order to find new cancer genes, two main mechanisms have been used : retrovirus and transposons. Retrovirus (e.g. murine leukemia virus (MuLV) and mouse mammary tumour virus (MMTV)) is a virus which its genome has a form of ribonucleic acid (RNA)and has the ability to convert its sequence into DNA by reverse transcription. Transposon (e.g. sleeping beauty and piggybac) is a DNA sequence that changes its position within the genome.

It is not known why this integration happens in the vicinity of a cancer gene. However, these mechanisms have integration biases [38].

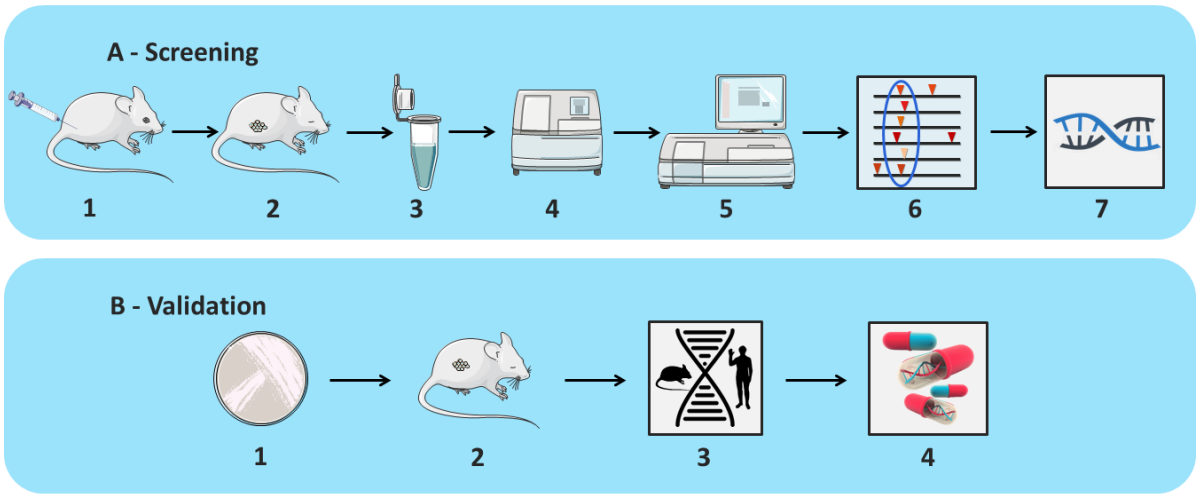


Figure 2.2: Outline for the cancer gene discovery using insertional mutagenesis (IM).
A- The mouse is infected with a retrovirus or transposon (1). After create a tumour (2), DNA is extracted (3), amplified -by polymerase chain reaction (PCR)- (4), sequenced and mapped (5). In order to find clusters of insertions some statistical and bioinformatics analysis is performed, also knows as CIS(6). Genes in the vinicity of CIS are potential cancer genes (7). **B-** After find new candidates to be a cancer genes, they must be validated. This validation consists in verify if the gene transform normal cells in cancer cells. It can be tested *in vitro* (1) and/or *in vivo*(2). If the transformation happens, a cross-species to find the orthologues and homologs genes in human is performed (3). The final step is create a drug which can correct the abnormal gene expression (4).

3

MACHINE LEARNING

3.1. LEARN FROM EXAMPLES

Machine Learning (ML) is a branch of computer science emerged from the study of artificial intelligence, pattern recognition and computational learning theory. This discipline is deployed in several fields, such as bioinformatics (e.g. evolution, systems biology, genomics and others [39]), medical diagnosis [40], computer vision (e.g. image recognition [41]), speech recognition [42], document classification (e.g. spam [43]), music [44], games (e.g. checkers [45]) and others.

The main goal of ML is to extract knowledge from experimental data, allowing the computer to make accurate predictions and decisions.

All ML problems start with a *dataset*, a collection of information. This information, also called *experiences* or *instances*, are individual and independent examples given to the learner, representing observations. Each experience is characterized by its values, representing a set of *features*. A feature (also known as *attribute*) is a measurement of something and can be nominal and numeric. Usually the dataset is defined as a matrix where the rows (m) are the instances collected, and the columns (n) are the features, representing the dimensionality of the data.

ML methods can be subdivided into two main groups based on the type of problem they can solve:

Supervised learning The learner gets a set of instances with their respective label, $D(X, Y)$, where D is the dataset, X is the set of vectors and Y the label. This method can be divided into two groups: *classification* and *regression*. The main difference between

these two analyses is the output type. In classification, the result is a discrete value, representing a class. In regression, the output is a continuous value and it depends on the independent variable given.

Unsupervised learning The learner gets a set of instances without labels, not being able to evaluate the method's error. The approaches used are essentially *clustering* and *dimensionality reduction*. The major difference between them is the way the reduction is done before their performance. In clustering, the number of experiences is reduced to generalize them. In dimensionality reduction, it is cutback the number of features, transforming them and reducing the dimensionality (preferably in two dimensions (2D) or three dimensions (3D)) to be easier to visualize.

However, there are more types of methods with more complex learning scenarios[3, 46].

3.1.1. CLASSIFICATION

Classification is used to identify in which set of categories a new experience can be labelled according to other experiences. The simplest classification problem is a binary classification. It creates a barrier (decision boundary) which separates the data in two different classes (Figure 3.1).

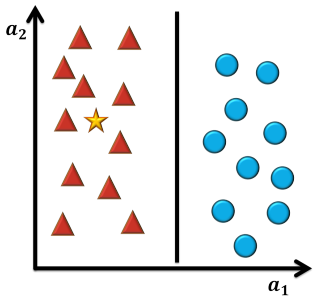


Figure 3.1: Binary classification.
Giving the data of two classes (triangles and circles)in two dimensions, where a_1 and a_2 represent two features, it is possible to separate both classes with a straight line (linear classifier). According to this decision boundary created, it is possible to classify the new experience (star). The new experience belongs to the class of triangles.

The decision boundary is created by an algorithm, named *classifier*. This function takes the unlabelled examples and maps them into labelled, using internal data structures. The learning task in classification problems is to construct classifiers which are able to classify unseen examples ($\hat{x}$) and give them a label ($\hat{y}$). A good classifier is the one who, given a set of experiences - *training set* -, to create the knowledge, is able to predict/classify new examples correctly. There are a large number of classifiers and each one can have different performances depending on the dataset. Six learning classifiers are described bellow:

Nearest Mean Classifier

The Nearest Mean Classifier (NMC) [47], also known as Minimum Distance Classifier, is a linear classifier. This classifier calculates the centre of the class. A new experience is classified according to the closest distance of all class centre (Figure 3.2).

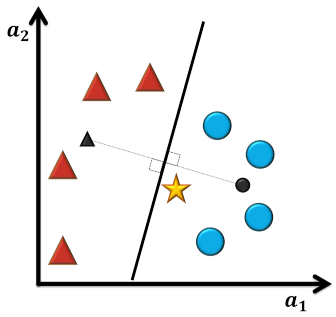


Figure 3.2: Example of Nearest Mean classification.
The centre of the classes are represented by the black triangle and circle. The classifier separates both classes creating a line equidistant to both centres. The test sample (star) should be classified as circle.

Giving a set of vectors representing the class y_1 (X^{y_1}), containing m samples with size n:

$$X^{y_1} = \{\vec{x}_1^{y_1}, \vec{x}_2^{y_1}, ..., \vec{x}_m^{y_1}\} \tag{3.1}$$

the centre of a class is determinated calculating the arithmetic mean ($\bar{X}$) of class's feature,

$$\bar{X}^{y_1} = \frac{1}{m^{y_1}} \sum_{i=1}^{m^{y_1}} \vec{x}_i^{y_1} \tag{3.2}$$

To classify a new experience $\hat{x}$, it is calculated the minimum distance d_E (Eucledian distance) between $\hat{x}$ and the centre of all classes ($\bar{X}_Y$).

$$\hat{y} = \underset{\{\bar{X}^{y_1}, \bar{X}^{y_2}, ..., \bar{X}^{y_z}\} \in \bar{X}_Y}{\arg\min} d_E(\hat{x}, \bar{X}_Y) \tag{3.3}$$

$$d_E(\vec{a}, \vec{b}) \equiv ||a - b|| \equiv \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{3.4}$$

k-Nearest Neighbour classifier

The k-Nearest Neighbour (kNN) classifier [48] classifies experiences based on closest training examples in the feature space. A new experience is classified by a majority

vote of the neighbourhood. The label more common of the k closest elements is the label of the new experience (Figure 3.3).

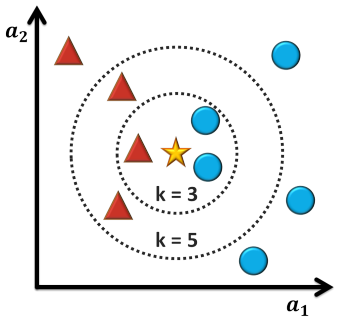


Figure 3.3: Example of k-Nearest Neighbour classification.
The test sample (star) should be classified either to the class of triangles or to the class of circles. If k = 3 (smaller dashed circle) it is assigned to the class of circles because there are 2 circles and only 1 triangle inside the inner circle. If k = 5 (bigger dashed circle) it is assigned to the class of triangles because there are 3 triangles and only 2 circles inside the inner circle.

Giving a training set D_t and a new experience $\hat{x}$, the method calculates the distance between the instance $\hat{x}$ and all training objects $(X, Y) \in D_t$, where X represents the set of vectors and Y its labels. Once all experiences are sorted by the closest distance, the new experience is classified based on the majority class of its k nearest neighbours:

$$\text{Majority voting: } \hat{y} = \underset{\{y_1, y_2, \ldots, y_z\} \in Y}{\arg\max} \sum_{i=1}^{k} I(y_i, Y), \tag{3.5}$$

where y_i are the labels of the nearest neighbours of class $\hat{x}$, k is the number of the neighbours, and $I(y_i, Y) = 1$ if $y_i = Y$ and $I(y_i, Y) = 0$ otherwise.

Support Vector Classifier

The Support Vector Machine (SVM) [49] is an algorithm which creates a hyperplane able to classify all training vectors in two classes. The best choice will be the hyperplane that leaves the maximum margin from both classes (Figure 3.4).

The formula for the output of a linear SVM is

$$\hat{y} = sgn(\vec{w} \cdot \vec{x} + b) \tag{3.6}$$

where $\vec{w}$ is the normal vector to the hyperplane, $\vec{x}$ is the input value and b the y-intercept. If $sgn(\vec{w} \cdot \vec{x} + b) < 0$, the new experience is classified by the class below the decision boundary, if $sgn(\vec{w} \cdot \vec{x} + b) > 0$ it is classified by the class above the boundary.

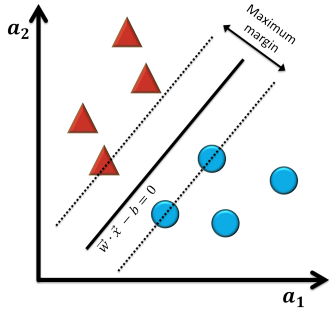


Figure 3.4: Example of Support Vector Machine classification in a linearly separable binary dataset.
The line is the hyperplan and the dashes lines are the margins. The main goal is to maximize the distance between margins. Samples on the margin are called the support vectors.

Naïve Bayes Classifier

The Naïve Bayes classifier [50, 51] is based on Bayes' theorem. This classifier assumes that the value of a feature is independent of the value of any other feature. This classifier learns the conditional probability (Equation 3.7) of each class label y_i given the attribute a_i:

$$P(y_i|a_i) = \frac{P(a_i|y_i)P(y_i)}{P(a_i)} \tag{3.7}$$

where $P(y_i|a_i)$ is the posterior probability of class (target) given predictor (attribute); $P(y_i)$ is the prior probability of class; $P(a_i|y_i)$ is the likelihood which is the probability of predictor given class; $P(a_i)$ is the evidence probability of predictor.

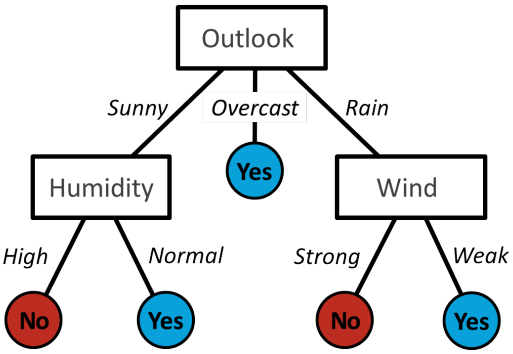
Classification is then done applying this probability of y_i given the particular instance of $\{a_1, a_2, ..., a_n\} \in A$. The predictable class is the one with the highest probability [52]

$$\hat{y} = \underset{y_i \in \{y_1, y_2, ..., y_z\}}{\operatorname{argmax}} P(y_i) \prod_{j=1}^{n} P(a_j|y_i) \tag{3.8}$$

Decision Tree Classifier

Decision tree [53] uses a tree structure. It breaks the dataset into smaller subsets until the result is a tree with a *decision node* or a *leaf node*. A decision node has two or more branches. A leaf node represents a classification. Decision tree requests several questions to attributes. Each answer will correspond to a branch. Once the decision tree is constructed, the classification is straightforward (Figure 3.5).

The simplest algorithm to construct decision trees is the Iterative Dichotomiser 3 (ID3) [53]. The major choice of ID3 algorithm is to find which attribute should be



Test data

Outlook	Temperature	Humidity	Wind	Play
Rain	Hot	Normal	Weak	$\hat{y}$=?

Figure 3.5: Example of a decision tree classification.
Nodes (rectangle) represent the features (outlook;humidity;wind), branches are the different answers for a feature and leaves (circle) are the output (yes/no). The test sample would be sorted downs the rightmost branch of the decision tree and would be classified as a positive instance (it is possible to play).

the root, the most appropriate to classify examples. This algorithm uses a statistical test - *Information Gain (IG)* - that measures how well a given attribute classifies experiences. ID3 uses this measure to select among the different attributes at each step wile growing the tree.

Before calculating *IG*, the variable *Entropy(H)* hve to be determined:

$$H(D_t) = \sum_{i=1}^{z} -P(y_i)\log_2 P(y_i) \quad , y_i \in Y \tag{3.9}$$

where D_t is the training set for which entropy is being calculated, Y the set of classes of D_t, z the number of different labels and $P(y_i)$ is the probability of y_i in Y. If $H(D_t) = 0$, the set D_t is perfectly classified.

IG is calculated according the following equation:

$$IG(D_t, A) = H(D_t) - \sum_{v \in V(A)} \frac{\#D_v}{\#D_t} H(D_v) \tag{3.10}$$

where $H(D_t)$ is the entropy of the training set D_t, $V(A)$ is the set of possible values for the attribute A, $\frac{\#D_v}{\#D_t}$ is the proportion of a value v and the size of the training set D_t and $H(D_v)$ is the entropy of the subset D_v.

Random Forest Classifier

The random forest classifier [54] uses the *bagging* method. Bagging is an approach

which creates additional data for training from the original dataset, creating several subsets, with random samples.

This classifier uses this approach to build a large collection of non-correlated trees. It selects a few combinations of samples with repetitions to create a decision tree. To predict a new element, all decision trees created are tested. Then, it counts how many time each class is predicted. The final result is the class with more votes.

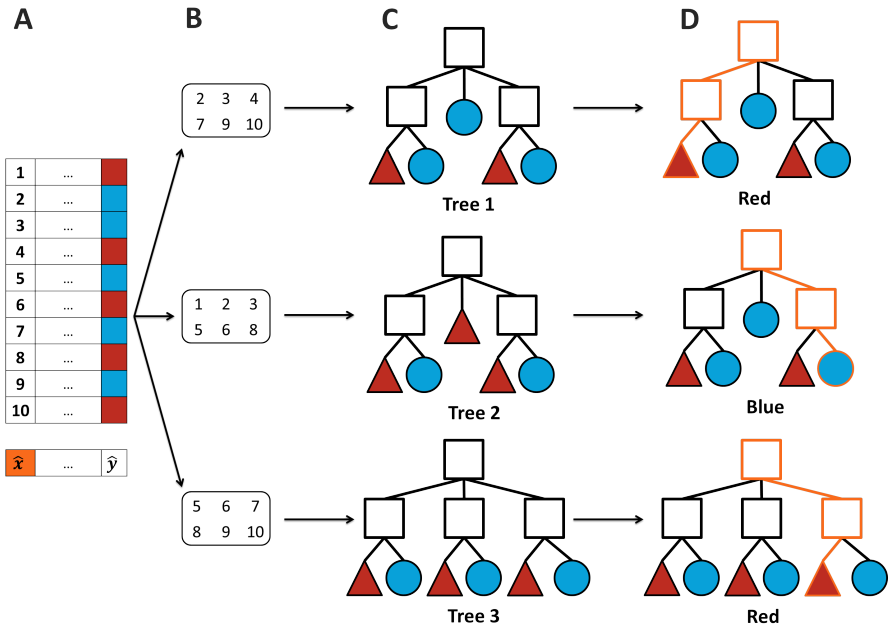


Figure 3.6: Example of a random forest classification.
A- The given dataset contains ten samples. Each sample has their respective label: blue or red. The main goal is to predict which colour corresponds the new example $\hat{x}$, in orange. **B-** Three subsets containing six samples from the initial dataset is creted. **C-** For each subset, a decision tree is created. **D-** In each tree, the new example $\hat{x}$ is predicted. The output of tree 1, tree 2 and tree 3 is red, blue and red, respectively. Counting the number of votes of each label, the final result of this classification is red.

It is important to understand that no classifier is 100% precise to solve all ML problem. The dataset also affects the classifier's performance. It also depends on the structure of the data (high/low bias and variances) and/or if a class has enough training experiences. A good way to find a classifier with a good performance is using *cross-validation*, testing their accuracy (see Section 3.2).

3.1.2. REGRESSION

Regression is used to find a predictive modelling which tries to find a relation between a dependent (x) and independent variable (y). The model (function) created should fit in

real data points. In contrast with classification problems (see Section 3.1.1), the output value, $\hat{y}$, is a continuous number.

There are several kinds of regression methods, but the simplest one is the linear model, represented by a linear equation $y = mx + b$ (Figure 3.7).

The model which has the *best fit* for a giving training data is calculated by minimizing the sum of the squares of the vertical distances from the data point to the line - *minimization of the sum of squared errors (SSE)* [55]:

$$SSE = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{3.11}$$

where m is the number of experiences, y_i is value of the dependent variable inx_i and $\hat{y}_i$ is the value of the dependent variable predicted by the model in x_i.

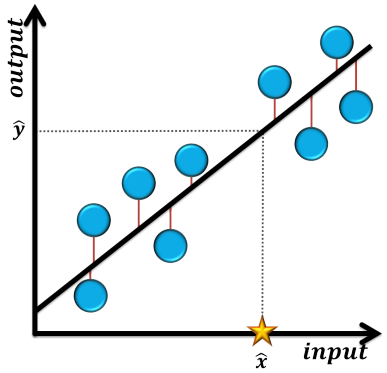


Figure 3.7: Example of linear regression.
The figure shows the linear regression line (black line), created from 9 training examples (circles). The connection between the data points and the point on the regression line (red lines) has the same x_i value, denoting the distance used to calculate the sum of squared errors.
The label of the new experience (star) is $\hat{y}$, $\hat{y} \in \mathbb{R}$.

3.1.3. CLUSTERING

Clustering analysis is used to group a set of experiences into subsets, differentiating each group (cluster) according to a certain criterion. Examples in the same cluster are more similar to each other compared with objects in other clusters. There are several clustering algorithms and, for that reason, it is not easy to have an exact definition of *cluster* [56, 57]. However, it is typically mentioned as a method to "group unlabelled data objects".

The main goal of this analysis is to understand how similar (or dissimilar) an individual experience is from other experiences. There are several different representations such as partitioned cluster and hierarchical cluster (Figure 3.8).

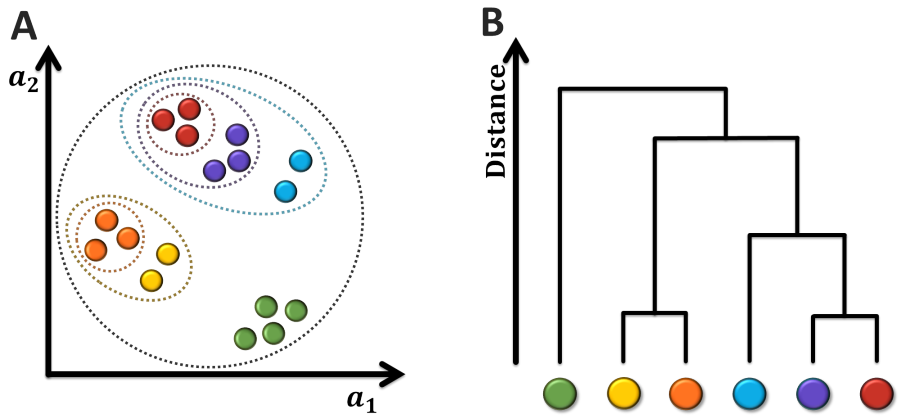


Figure 3.8: Cluster analysis.
A- Partitional clustering. The experiences are projected in a 2 dimensional plane. It is possible to group some examples according to their similarity of the features a_1 *and* a_2. **B-** Hierarchical clustering. Taking the clusters of **A**, it is possible to calculate the distances between them and represent it in a diagram tree or dendogram. Each circle represents an experience and colours are used to distinguish clusters.

3.1.4. DIMENSIONALITY REDUCTION

In Machine Learning problems, most of the data has a high dimension, in other words, a large number of features (n). In several domains it is important to visualize the data, but, with a high number of featuresm it can be difficult to extract information. For that reason, before analyse the data, a *dimensionality reduction* should be performed. This process takes the initial data and transforms into a lower-dimensional representation, preserving some properties of the initial form. The dimensionality reduction can be divided into feature selection and feature extraction:

1. **Feature selection**
   The feature space is reduced by selecting a subset of revelant features from the original data.

2. **Feature extraction**
   The feature space from the original data is reduced through some functional mapping. After feature extraction, the features are transformed and reduced. The new attributes are $A' = \{a'_1, a'_2, \ldots, a'_{nr}\}$, with $(nr < n)$ and $A' = F(A)$, where F is a mapping function, which transforms the attribute A into A', nr is the number of features after reduction. There are several feature extraction algorithms, but it will be present the Principal Component Analysis (PCA) and the t-distributed stochastic neighbor embedding (t-SNE):

   Principal Component Analysis (PCA)
   The central idea of PCA [58–60] is to convert the data using an orthogonal trans-

formation. It will transform the data into a set of uncorrelated linear variables - *principal components*. The principal components are ordered according the degree's variance. The first principal components contain most of the variation present in all of the original variables. The succeeding components have the highest variance compared to the preceding components (Figure 3.9 B)

t-distributed stochastic neighbor embedding (t-SNE)

t-SNE [61] is a nonlinear dimensionality reduction method. It is well suited to reduce high-dimensional data into the space of two or three dimensions. This analysis minimizes the divergence between two distributions: construct a distribution that measures pairwise similarities, where similar samples have a high probability of being selected; and also construct a distribution that measures pairwise similarities of the corresponding low-dimensional maps (Figure 3.9 C).

Given that PCA and t-SNE are unsupervised learning, the labels of the data are not used in the transformation. However, they are used to colour intermediate plots.

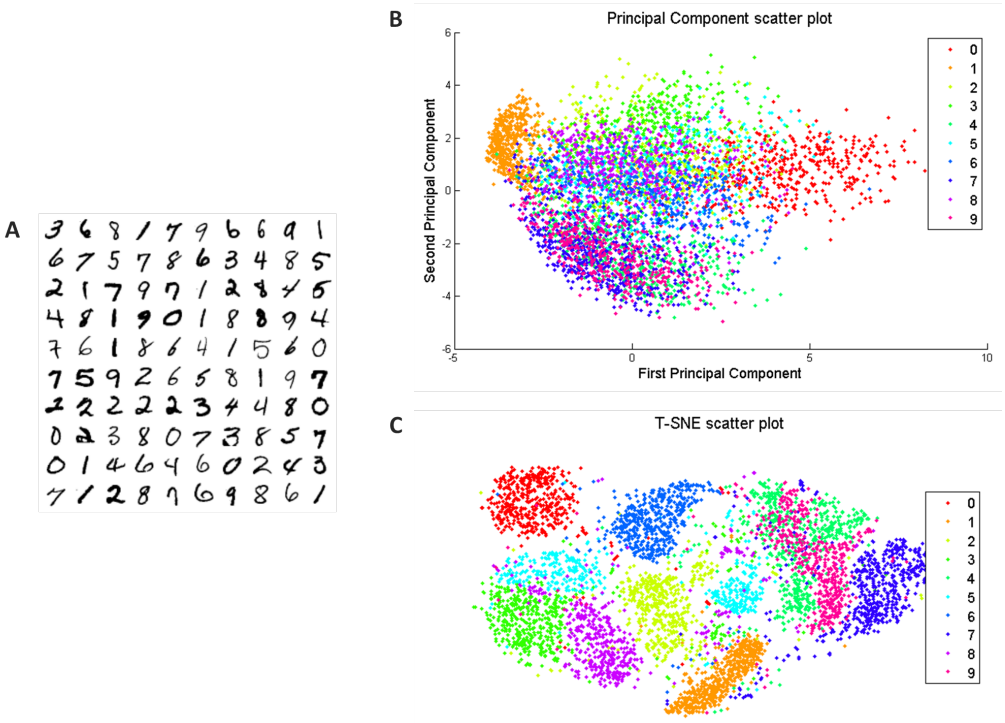


Figure 3.9: Visualization of 2,000 samples of the Mixed National Institute of Standards and Technology (MNIST) dataset using PCA and t-SNE.
A- The MNIST dataset contains information of handwritten digits (from LeCun, Bottou, Bengio, and Haffner (1998) [62]). PCA (**B**) and t-SNE (**C**) are dimensionality reduction algorithms, preserving the proprieties of the original dataset and allowing the visualization of the data. The figure shows that t-SNE has a better visualization compared to PCA.

3.2. CROSS-VALIDATION

Cross-validation (CV) is a statistical method, used in preticting problems, to evaluate the accuracy (or error) of a model, being able to evaluate learning algorithms [63].

To perform this analysis, the data is splitted into two groups: the *train set*, used to learn a model; and the *test set*, used to validate the model.

There are a few different types of CV, but the most used is the K-fold cross-validation [64]. In K-fold cross-validation the data is subdivided into K identical sized folds. For each K an iteration is performed: a different fold is used for a validation and the $K-1$ folds to learn. Each iteration has an error as an output. After all iterations, it is possible to calculate the average error rate of the model, giving an idea of how well the model generalizes (Figure 3.10).

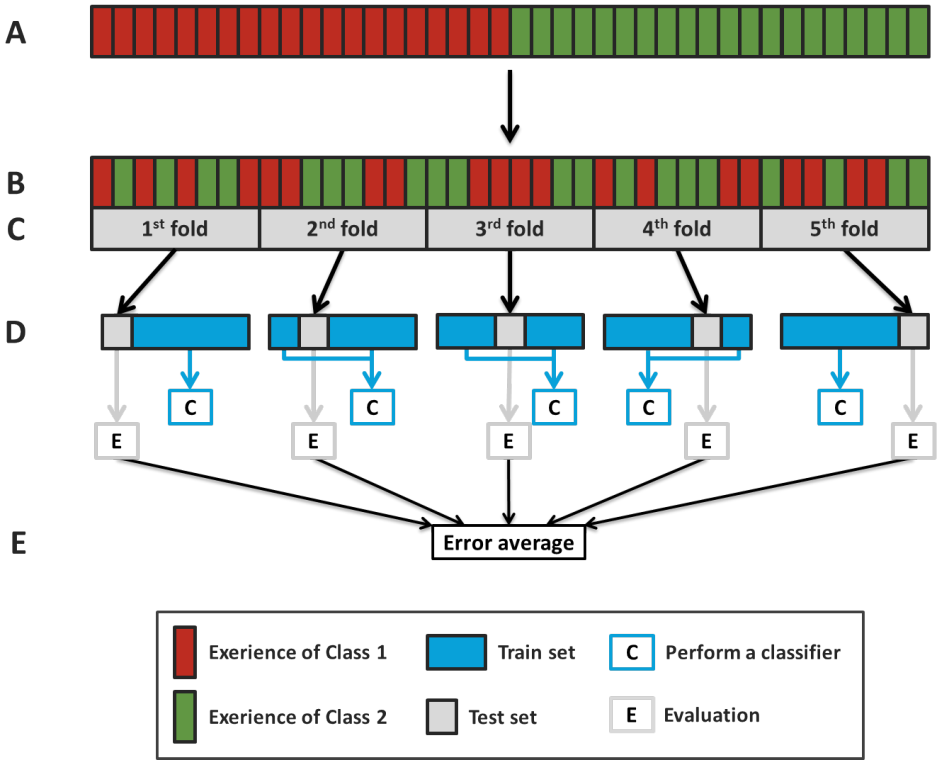


Figure 3.10: Representation of K-fold cross-validation.
A- Dataset contains experiences of two classes. **B-** Data is reshuffled randomly to reduce the bias. **C-** Data is subdivided into five identical sized subsets ($K = 5$). **D-** From the five folds created, four are used to train the model and the last fold for evaluation. **E-** The output is the average error rate of a classifier, giving an idea of how well is the classifier's performance. In order to reduce the error rate, this process can be repeated, giving a more accurate average of each evaluation. In each repetition, the data is reorganized (**B**).

The number of folds (K) to use is arbitrary, but there are some points to take into account: if a large value is used, the bias of the true error rate estimator will be small, but the variance of the true error rate will be large and it will take too many time, due the low number of experiments in each fold; If a small number of folds is used, the computation

time is reduced, the variance of the estimator will be small, but the bias of the estimator will be large. A common choice for this method is use $K = 10$.

The output of each iteration is the estimated accuracy of the model. The accuracy of a classifier C is the probability of classifying correctly a random experience, *i.e.*, $acc = P(C(\hat{x}) = \hat{y})$, where $\hat{x}$ is the experience and $\hat{y}$ its class.

In CV, the accuracy (acc) corresponds to the number of correct classifications, divided by the number of instances in the dataset [64]:

$$acc_{CV} = \frac{1}{m} \sum_{(\hat{y}, y_i) \in D_t} I(C(D_t, \hat{y}), y_i) \tag{3.12}$$

where m is the number of instances of the training set D_t, $C(D_t, y_u)$ is the mapping function of the classifier C in the train set (D_t), having $\hat{y}$ as a result, and I is an indicator function where $I(a, b) = 1$ if $a = b$ and 0 otherwise.

The error (ERR) of a model can be calculated by:

$$ERR = 1 - acc \tag{3.13}$$

Another way to evaluate the viability of a model is using the area under the curve (AUC) of a receiver operating characteristic (ROC) curve [65]. Considering a two-class classification problem, in which the outcomes are $presence$ or $absences$ of a disease, it can have four possible solutions (Table 3.1): samples carrying the disease and the model can classify correctly its presence (True Positive (TP)), however, sometimes can happen to be classified as healthy (False Negative (FN)). On the other hand, some samples without the disease will be correctly classified as negative (True Negative (TN)), but some cases without the disease will be classified as positive (False Positive (FP)).

The ROC curve is created by plotting the False Positive Rate (or $1 - Specificity$) against the True Positive Rate (or $Sensitivity$):

$$sensitivity = \frac{TP}{TP + FN} \qquad specificity = \frac{TN}{TN + FP} \tag{3.14}$$

The curve always goes through two points: (0,0) and (1,1) (Figure 3.11). The model is considered better than other if the AUC is greater. If the AUC is equal to 1, it means that the test is 100% accurate because both specificity and sensitivity are 1, without false negative and false positive values. On the other hand, if a test cannot assess between correct and incorrect, the curve will correspond to a diagonal, where its AUC is equal to 0.5. The typical AUC of a ROC curve is between 0.5 and 1.

Table 3.1: Confusion matrix used to tabulate the predictive capacity of presence/absence models. It can have four different outcomes: True Positive (TP) - presence observed and predicted by model; False Positive (FP) - absence observed but predicted as present; False Negative (FN) - presence observed but predicted as absent; True Negative (TN) - absence observed and predicted by model.

		Predicted	
		Presence	Observed
Actual	Presence	TP	FN
	Absence	FP	TN

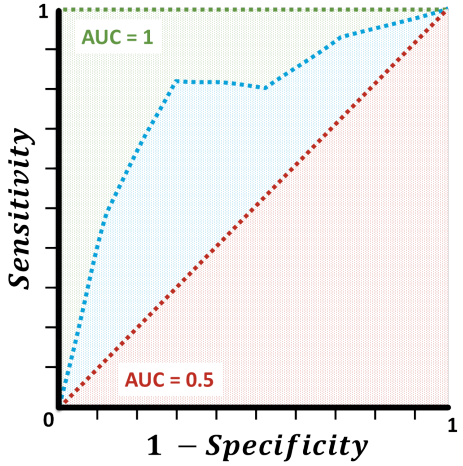


Figure 3.11: Representation of three ROC curves. The green curve ($AUC = 1$) represents the best model, while the red curve ($AUC = 0.5$) represents the worst one. The blue curve is a positive predictive model.

The error (ERR) of a model can be calculated by:

$$ERR = 1 - AUC \tag{3.15}$$

3.3. DISSIMILARITY REPRESENTATION

In many cases it is not easy to evaluate a dataset and compare its samples. It can be convenient to understand how different two samples are, that is, the *distance* (or *dissimilarity*) between them. Considering $d(a,b)$ the dissimilarity of the sample a from b, then

$$\begin{aligned} d(a,b) &> 0 && \text{if } a \neq b \\ d(a,b) &= 0 && \text{if } a = b \\ d(a,b) &= d(b,a) \\ d(a,b) &\leq d(a,c) + d(b,c) \end{aligned} \tag{3.16}$$

If the dissimilarity measures satisfy the four conditions above, the dissimilarity measure is a *metric* and the term *distance* is usually used [66].

Compare all samples of a dataset will generate a *distance matrix* [67, 68]. Here, a distance matrix is considered as a 2D array containing the distances, taken pairwise, between the samples of a dataset.

Matrix 3.17 represents an example of an array M($m \times n$), m rows and n columns, and it is filled with Boolean data, $B = \{0, 1\}$. Each row is a vector ($\vec{x}_i$) and each column an attribute (a_j).

$$\mathrm{M} = \begin{array}{c} \\ \vec{x}_1 \\ \vec{x}_2 \\ \vec{x}_3 \\ \vec{x}_4 \\ \vdots \\ \vec{x}_m \end{array} \begin{array}{c} \begin{array}{cccccccc} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & \ldots & a_n \end{array} \\ \begin{bmatrix} B & B & B & B & B & B & \ldots & B \\ B & B & B & B & B & B & \ldots & B \\ B & B & B & B & B & B & \ldots & B \\ B & B & B & B & B & B & \ldots & B \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ B & B & B & B & B & B & \ldots & B \end{bmatrix} \end{array} \tag{3.17}$$

Calculating the distance d through the matrix M, will map the distances between all samples of the dataset, creating a distance matrix (Matrix 3.18). The distance matrix is a square matrix with size $m \times m$, symmetric, filled with non-negative elements and the diagonal elements are equal to zero. These proprieties are justified by the equations 3.16.

$$d(\mathrm{M}) = \begin{bmatrix} 0 & d(\vec{x}_1, \vec{x}_2) & d(\vec{x}_1, \vec{x}_3) & d(\vec{x}_1, \vec{x}_4) & \cdots & d(\vec{x}_1, \vec{x}_m) \\ d(\vec{x}_2, \vec{x}_1) & 0 & d(\vec{x}_2, \vec{x}_3) & d(\vec{x}_2, \vec{x}_4) & \cdots & d(\vec{x}_2, \vec{x}_m) \\ d(\vec{x}_3, \vec{x}_1) & d(\vec{x}_3, \vec{x}_2) & 0 & d(\vec{x}_3, \vec{x}_4) & \cdots & d(\vec{x}_3, \vec{x}_m) \\ d(\vec{x}_4, \vec{x}_1) & d(\vec{x}_4, \vec{x}_2) & d(\vec{x}_4, \vec{x}_3) & 0 & \cdots & d(\vec{x}_4, \vec{x}_m) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d(\vec{x}_m, \vec{x}_1) & d(\vec{x}_m, \vec{x}_2) & d(\vec{x}_m, \vec{x}_3) & d(\vec{x}_m, \vec{x}_4) & \cdots & 0 \end{bmatrix} \tag{3.18}$$

There are many different ways to measure dissimilarity and, for that reason, there are many different dissimilarity transformations. It depends upon the application involved. For vectors of binary data, $\vec{x}_i$ and $\vec{x}_j$, these may be expressed in terms of the number of a, b, c and d where

a is equal to the number of occurrences of $\vec{x}_i = 1$ and $\vec{x}_j = 1$

b is equal to the number of occurrences of $\vec{x}_i = 0$ and $\vec{x}_j = 1$

c is equal to the number of occurrences of $\vec{x}_i = 1$ and $\vec{x}_j = 0$

d is equal to the number of occurrences of $\vec{x}_i = 0$ and $\vec{x}_j = 0$

This is summarised in Table 3.2.

Table 3.2: Co-occurrence table for binary variables

		$\vec{x}_i$	
		1	0
$\vec{x}_j$	1	a	b
	0	c	d

Two metrics often used are presented to map binary data into distances matrix:

1. **Hamming distance**

   The Hamming dissimilarity [69] is defined by the ratio of mismatches among their pairs of variables:

$$d_H = \frac{\#(\vec{x}_i \neq \vec{x}_j)}{\#[(\vec{x}_i \neq \vec{x}_j) \cup (\vec{x}_i = \vec{x}_j)]} \equiv \frac{b+c}{a+b+c+d} \tag{3.19}$$

2. **Jaccard distance**

   The Jaccard dissimilarity[70] is defined by the ratio of mismatches among the non-zeros's pairs:

$$d_J = \frac{\#[(\vec{x}_i \neq \vec{x}_j)}{\#[(\vec{x}_i \neq 0) \cup (\vec{x}_j \neq 0)]} \equiv \frac{b+c}{a+b+c} \tag{3.20}$$

Equation A.1, in Appendix A.1, shows 9 examples to compare both metrics.

3.4. FEATURE RANKING

In Machine Learning, feature ranking is used to sort features, by relevance, for a certain class in a two class task. Different methods have been developed depending on the application [71]. However, this can bring some issues. Different methods will generate a different feature ranking of the same data.

A recent study [72] compares the three ranking algorithms for binary features to understand which one generates the most 'correct' ranking. Using five artificial data and four real-life data they concluded that the *diff-criterion* algorithm got the most correct performance.

Diff-criterion [73] uses a distance between probability distributions of two classes. It estimates the importance of the i^{th} feature as:

$$\vec{R} = p(a_i = 1|y_1) - p(a_i = 1|y_2) \tag{3.21}$$

where $p(a_i = 1|y_1)$ and $p(a_i = 1|y_2)$ are the probability of a feature has a 1 in the classes y_1 and y_2. $\overrightarrow{R}$ is a vector containing the scores of a feature a_i. Each score is a value between -1 and 1. The higher the score, the greater importance. If a score is zero, it means the feature has the same probability of belonging in both classes. Sorting $\overrightarrow{R}$, it is possible to have the attributes sorted according to that parameter.

4

DATA

4.1. DATA GENERATION

The following work was developed using an exclusive data collection, compiled from several studies (Table 4.1). They correspond to a compendium of IM screens in mice. The data from each study are available online and can be downloaded.

Each study uses several samples of tumour development. All of them were infected with an integration element (e.g. a retrovirus or a transposon). After that, insertions are identified and mapped. For each gene, two windows with 10 kilobase (kb), up- and downstream of its location, were created. *Window space* is the name given to the distance between the first bp of a window upstream of the gene and the last bp of a window downstream of the gene (comprising the gene). It was verified, in all window spaces, whether it is carried an insertion or not. This results in a Boolean matrix: if an insertion is included in a window space, the gene will contain a 1, otherwise, the gene has a 0. All the information is stored in a *.csv* file (Figure 4.1).

4.2. DESCRIPTION

This dataset contains information collected across 54 studies, obtained from 38 papers (Table 4.1). This compilation contains 7,037 samples of IM data organized in 14 types of cancer. To the best of my knowledge, this is the first analysis, using DNA integration, to span an extensive number of independent tumours.

Each study contained between 17 and 3853 samples. They are related to one type of cancer, resulting in 13 specific ones and in one additional type labelled “Various”, which

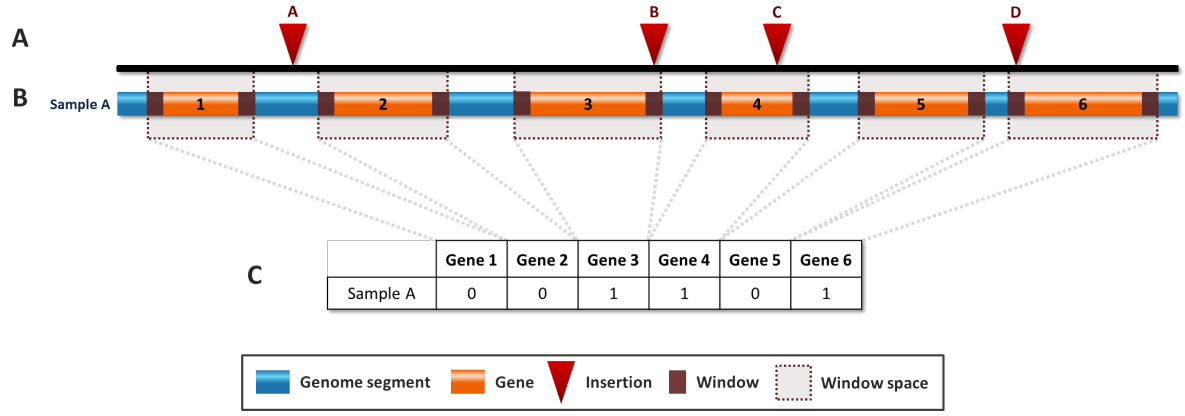


Figure 4.1: Organization of data generated.
A- Each sample represents the genome of a cancer cell in the mouse. The mouse is infected by an integrating element which inserts, randomly, pieces of DNA in the genome. After mapping the insertions, it is possible to identify their location. **B-** For each gene, two windows of 10 kb were created, located up- and downstream of the gene. In the area covered between windows - window space -, it was verified whether an insertion was integrated or not. **C-** The previous analysis was stored in a matrix, where for each sample it is identified which genes have an insertion in its vicinity. Insertion A is not covered by any windows. Insertion B (downstream of the gene), insertion C (within the gene) and insertion D (upstream of the gene) are captured by the window space of gene 3, gene 4 and gene 6, respectively. For this reason, entries for gene 1, gene 2 and gene 5 are 0, and gene 3, gene 4 and gene 6 are 1.

contain information of more than one cancer type. All types of cancer are described below:

Basal cell carcinoma

Basal cell carcinoma (BCC) it is the most common type of skin in cancer (80%)[2] and one of the most common type of cancer in humans. It is typically developed on areas that have been exposed in the sun. It growths slowly and spreads to the nearest tissues, but it is rare to spread to other body parts.

Colorectal

Colorectal cancer starts in the colon or the rectum (part of the large intestine). It is the third most commonly diagnosed cancer in males and the second in females [2].

Glioblastoma

Glioblastoma multiforme (GBM) the most common type of cancer in the nervous system. It is formed from glial tissues of brain and spinal cord.

Hematopoietic

Hematopoietic stem cells (HSC) can develop all types of blood cells, producing an enormous number of blood cells every day.

[2]http://www.cancer.org/cancer/skincancer-basalandsquamouscell/detailedguide/skin-cancer-basal-and-squamous-cell-what-is-basal-and-squamous-cell, accessed: July 2015

Hepatocellular carcinoma

Hepatocellular carcinoma (HCC) is the most common form of liver cancer. It is the second leading cause of cancer death in males [2].

Lymphoma

Lymphoma is a cancer which starts in the lymphoma system, a part of the immune system.

Malignant peripheral nerve sheath

Malignant peripheral nerve sheath tumour (MPNST) is a variety of soft tissue tumours. It is a rare tumour and appears in a neuron cell, the Schwann cells.

Mammary

Mammary cancer, also known as breast cancer in Human, is originated in the mammary gland. It is the most common cause of death in females.

Medulloblastoma

Medulloblastoma is the most common paediatric primary brain tumour. It can begin in the lower part of the brain and spread to the spine or other part of the body.

Pancreatic

Pancreatic cancer starts in the pancreas. It is one of the most lethal type of cancer because usually is only diagnosed in advanced stages [74].

Sarcoma

Sarcoma is a type of cancer that begins in bone or in the soft tissues of the body (e.g. muscle, fibrous tissue, cartilage, *etc*).

Squamos cell carcinoma

Squamos cell carcinoma (SCC) is the second most common skin cancer, after BCC[3]. Like BCC it also develops on sun-exposed areas. It growth more likely into deeper layers of skin and are it is more frequent to spread to other body parts, comparing with BCC, but it is still uncommon.

T-cell acute lymphoblastic leukaemia

T-cell acute lymphoblastic leukaemia (T-ALL) starts in one of the lymphocytes' category: T-cell. It is a type of white blood, present in the immune system.

[3]http://www.skincancer.org/skin-cancer-information/squamous-cell-carcinoma, accessed: July 2015

In general, each tumour type has a few thousands samples (7037 in total) and all of them refer to the mouse genome, representing 22019 genes.

Commonly the genome has only a few insertions. Genes with insertions in their vicinity represent 0.0759% of the entire data.

4.3. PRE-PROCESSING

This data contains information about 13 different tumour types and an extra containing analysis of several cancer types, named "various" [79, 82, 83]. For the purpose of the present work, this last group was too ambiguous, not giving information about a particular tumour type. For this reason, this set, representing 86 samples, was removed.

In a first step, the distributions of insertions per sample and gene were analysed (Figures 4.2a and 4.2b). In general, it is shown that, in both situations, it is more frequent to have a few insertions. In fact, more than 3,000 samples have less than 3 insertions and more than 10,000 genes have less than 4 insertions.

The insertion can happen in the entire genome. However, it does not mean it is close to a gene. Therefore, the window space may not catch the integration. In order to have more informative samples, the median of insertions' frequency was used as a threshold. Samples which have less than 4 insertions were removed. In some tumour types, this elimination results in a loss of more than 60% of samples, or even the total loss of samples (Table 4.2). In total, this threshold excludes 3,244 samples.

After the samples' removal, some tissues had just a few numbers of examples. It is not valid to perform an analysis between two classes which have a large difference in numbers of samples (e.g. compare lymphoma versus SCC, with 98.8% less samples). To have a statistically significant analysis, all tumour types which have less than 10% of lymphoma's sample size were excluded. In other words, all cancer types which have less than 130 samples, after the threshold process, were removed. They are: basal and squamos cell carcinoma;glioblastoma; mammary; pancreatic; sarcoma and squamos cell carcinoma (corresponding, together, 293 samples).

If a gene has a few insertions, it is not too informative. It means that some genes are not involved in the tumourogenesis' process. It is more interesting if a gene has a lot of insertions in the vicinity of a gene in independent tumours (common insertion site (CIS)). Similarly to the samples' analysis, a threshold it was used to remove that genes that are not so interesting. This threshold corresponds to the median of the frequencies. Genes which have less than 5 insertions were removed. This removal corresponds to 14,268 genes.

After all this cut-off, the data used in the following project corresponds to a Boolean ma-

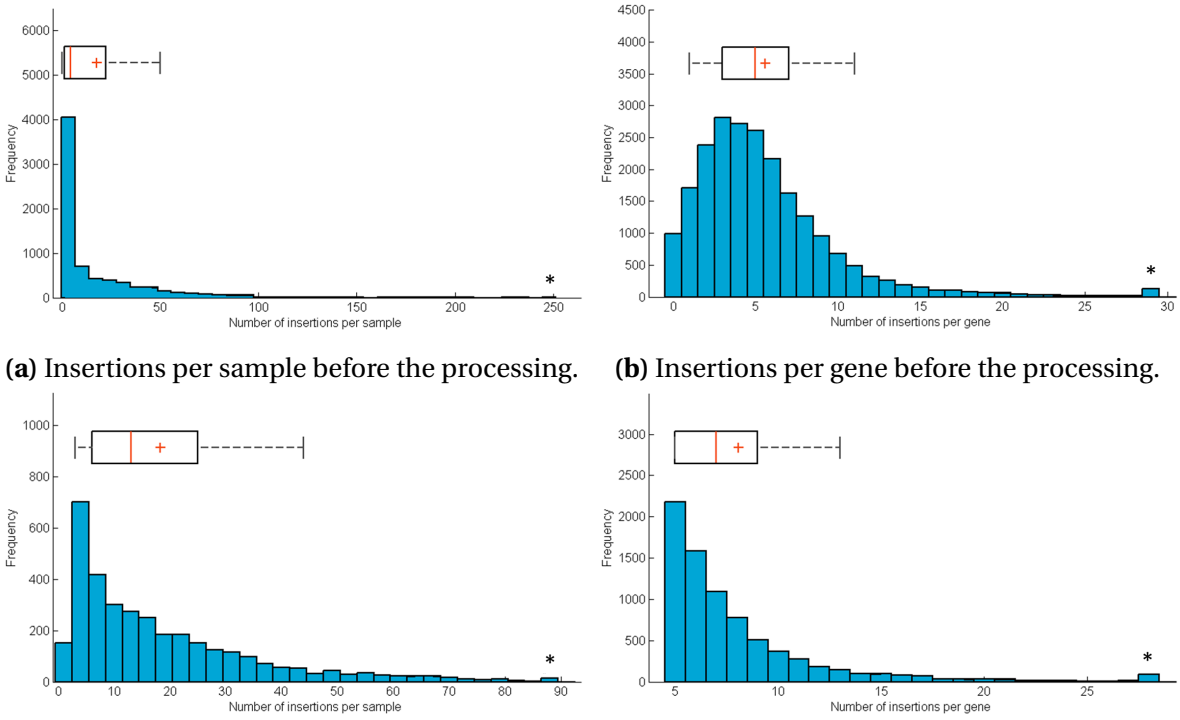
Table 4.1: List of studies collected for this project regarding to insertional mutagenesis screens. It contains 14 different tumour types, with several samples, totalling 7037.

	Study name	Tumour type	Number of Samples	Reference
1	BARD_NATURE-GENETICS_2014_ALL	Hepatocellular carcinoma	250	[75]
2	BENDER_CANCER-RESEARCH_2009_BEN	Glioblastoma	21	[76]
3	BERQUAM-VRIEZE_BLOOD_2011_CD4	T-ALL	38	[77]
4	BERQUAM-VRIEZE_BLOOD_2011_LCK	T-ALL	27	[77]
5	BERQUAM-VRIEZE_BLOOD_2011_VAV	T-ALL	36	[77]
6	CESANA_MOL-THERAPY_2014_ALL	Hematopoietic	277	[78]
7	COLLIER_CANCER-RESEARCH_2009_LYM-LEU	Various	59	[79]
8	COLLIER_NATURE_2005_ALL	Sarcoma	28	[80]
9	DUPUY_CANCER-RESEARCH_2009_HCC	Hepatocellular carcinoma	11	[81]
10	DUPUY_CANCER-RESEARCH_2009_SCC	Squamos cell carcinoma	17	[81]
11	DUPUY_NATURE_2005_ALL	Various	16	[82]
12	FRIEDEL_PLOS-ONE_2013_ALL	Various	11	[83]
13	GENOVESI_PNAS_2013_MB	Medulloblastoma	85	[84]
14	HUSER_PLOS-GENETICS_2014_GIM1	Lymphoma	28	[85]
15	KENG_HEPATOLOGY_2013_COMB	Hepatocellular carcinoma	162	[86]
16	KENG_NATURE-BIOTECHNOLOGY_2009_HCC	Hepatocellular carcinoma	69	[87]
17	KOOL_CANCER-RESEARCH_2010_CDK	Lymphoma	1354	[88]
18	KOSO_CANCER-RESEARCH_2014_P53	Medulloblastoma	27	[89]
19	KOSO_CANCER-RESEARCH_2014_WT	Medulloblastoma	17	[89]
20	KOSO_PNAS_2012_CELL	Glioblastoma	26	[90]
21	KOSO_PNAS_2012_TUMOUR	Glioblastoma	70	[90]
22	KOUDIJS_GENOME-RESEARCH_2011_MULV	Mammary	48	[91]
23	KOUDIJS_GENOME-RESEARCH_2011_SB	Lymphoma	379	[91]
24	LATOWSKA_ANC_2013_MB	Medulloblastoma	41	[92]
25	MANN-K_PNAS_2012_KRAS	Pancreatic	21	[93]
26	MARCH_NATURE-GENETICS_2011_ALL	Colorectal	445	[94]
27	ODONNELL_PNAS_2012_ALL	Hepatocellular carcinoma	24	[95]
28	PEREZ-MANCERA_NATURE_2012_SB10	Pancreatic	58	[96]
29	PEREZ-MANCERA_NATURE_2012_SB13	Pancreatic	197	[96]
30	QUINTANA_INVESTIGATIVE-DERMATOLOGY_2013_SB11	Basal and Squamos cell carcinoma	75	[97]
31	RAD_SCIENCE_2010_ALL	Hematopoietic	91	[98]
32	RAHRMAN_NATURE-GENETICS_2013_NF	Malignant peripheral nerve sheat	267	[99]
33	RAHRMAN_NATURE-GENETICS_2013_PNST	Malignant peripheral nerve sheat	100	[99]
34	RANZANI_NATURE-METHODS_2013_ALL	Hepatocellular carcinoma	30	[100]
35	STARR_PNAS_2011_ALL	Colorectal	96	[101]
36	STARR_SCIENCE_2009_DATASET1	Colorectal	42	[102]
37	STARR_SCIENCE_2009_DATASET2	Colorectal	93	[102]
38	THEODOROU_NATURE-GENETICS_2007_ALL	Mamary	136	[103]
39	UREN_CELL_2008_P19KO	Lymphoma	617	[104]
40	UREN_CELL_2008_P53KO	Lymphoma	326	[104]
41	UREN_CELL_2008_WT	Lymphoma	454	[104]
42	VAN-DER-WEYDEN_BLOOD_2011_BCP-ALL	Lymphoma	15	[105]
43	VAN-DER-WEYDEN_BLOOD_2011_T-ALL	Lymphoma	19	[105]
44	VAN-DER-WEYDEN_CANCER-RESEARCH_2012_KO	Lymphoma	109	[106]
45	VAN-DER-WEYDEN_IJCR_2012_KO	Lymphoma	92	[107]
46	VAN-DER-WEYDEN_IJCR_2012_POOLED	Lymphoma	126	[107]
47	VAN-DER-WEYDEN_ONCOGENE_2013_HET	Lymphoma	116	[108]
48	VAN-DER-WEYDEN_ONCOGENE_2013_HOM	Lymphoma	9	[108]
49	VASSILIOU_NATURE-GENETICS_2011_NPM1C	Lymphoma	85	[109]
50	VASSILIOU_NATURE-GENETICS_2011_NPM1WT	Lymphoma	30	[109]
51	WONG_NATURE-GENETICS_2014_CUX1	Lymphoma	70	[110]
52	WU_NATURE_2013_PTCH	Medulloblastoma	140	[111]
53	WU_NATURE_2013_TP53	Medulloblastoma	33	[111]
54	ZANESI_BLOOD_2013_CD19-CRE	Lymphoma	24	[112]

Table 4.2: Samples' size reduction of each tumour types.
$^{(a)}$ Number of samples of each tumour type before the data treatment.
$^{(b)}$ Number of samples of each tumour type after the data treatment.
⋆ Type of tumour discarded.

	Tumour type	Number of Samples $^{(a)}$	Number of Samples $^{(b)}$	Removal percentage (%)
⋆	Basal and squamous cell carcinoma	75	73	2.67
	Colorectal	676	623	7.84
⋆	Glioblastoma	117	95	18.80
	Hematopoietic	368	140	61.96
	Hepatocellular carcinoma	546	453	17.03
	Lymphoma	3853	1308	66.05
	Malignant peripheral nerve sheath	367	329	10.35
⋆	Mammary	184	24	86.96
	Medulloblastoma	343	333	2.92
	Pancreatic	276	228	17.39
⋆	Sarcoma	28	0	100.00
⋆	Squamous cell carcinoma	17	16	5.88
⋆	T-ALL	101	85	15.84

trix of 3,414 objects (samples) by 7,751 features (genes) organized into 7 classes (colorectal, hematopoietic, hepatocellular carcinoma, lymphoma, malignant peripheral nerve sheath, medulloblastoma and pancreatic). Figures 4.2c and 4.2d shows the insertions' frequency after this pre-processing.



(a) Insertions per sample before the processing.

(b) Insertions per gene before the processing.

(c) Insertions per sample after the processing.

(d) Insertions per gene after the processing.

Figure 4.2: Distribution of insertions' frequency represented in histogram and boxplot.
a- Number of insertions per sample before the processing: Data contains 7037 values; the median is 4; the mean is 17.51; the * means that there are 12 points bigger than 244. **b-** Number of insertions per gene before the processing: Data contains 22019 values; the median is 5; the mean is 5.9; the * means that there are 130 points bigger than 28. **c-** Number of of insertions per sample after the processing: Data contains 3414 values; the median is 13; the mean is 18.34; the * means that there are 13 points bigger than 86. **d-** Number of insertions per gene after the processing: Data contains 7751 values; the median is 7; the mean is 8.08; the * means that there are 91 points bigger than 27.5.
+ - Mean of the values. | - Median of the values.

5

METHODOLOGY

For the several steps of the work, Matrix laboratory (MATLAB)[113] was used. It is a high-performance language for technical computing, integrating computation, visualization and programming.

5.1. DATA

The pre-processing of the data was done using MATLAB. The data are organized in a matrix, containing the information of 3,414 samples over 7,751 genes (see Section 4.3). Each sample is characterized by one label, representing the tumour type (colorectal, HSC, HCC, lymphoma, MPNST, medulloblastoma and pancreatic).

5.2. DATA TRANSFORMATION

In order to understand differences between samples, the distance method may be performed. Comparing two samples (two boolean vectors), three different situations can occur: a gene does not have insertions in both samples ($match_{0-0}$), a gene have insertion in both samples ($match_{1-1}$), or a gene have one insertion in only one of the samples ($mismatch$):

$$\begin{matrix} \vec{x}_1 \\ \vec{x}_2 \end{matrix} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{5.1}$$

Analysing the samples $\vec{x}_1$ and $\vec{x}_2$ of Matrix 5.1, it has six $match_{0-0}$, one $match_{1-1}$ and three $mismatch$.

Since the integrations affect only a few genes in the whole genome it is important to focus on the differences between them. Four different metrics were applied: two known from the literature (Hamming [69] and Jaccard distance [70]) and two new metrics were developed and implemented (gene dependent method (GDM) and gene independent method (GIM)).

The Hamming distance takes into account the mismatch genes and compares with all other situations. On the other hand, the Jaccard distance also takes into account the mismatch genes, but it ignores the $match_{0-0}$. This can be an advantage because this metric will take in consideration all genes that have an insertion at least in one sample. The Hamming and Jaccard distance were calculated using the function *pdist* from the Statistics and Machine Learning Toolbox of MATLAB.

These two metrics do not have in consideration the biological meaning of the data. To take into account this factor, two new metrics were developed - GDM and GIM. Both metrics are described in the next two sections.

5.2.1. GENE DEPENDENT METHOD

The previous methods consider all genes with the same weight, but it could be relevant distinguish them. For example, having three samples containing information about two genes: the first sample is the control, without mutations; the second has a mutated gene implied in the development of a disease; and the third one has a mutation in a gene which is not involved in the disease phenotype (Matrix 5.2)

$$\begin{array}{c} \\ Sample_0 \\ Sample_1 \\ Sample_2 \end{array} \begin{array}{c} \begin{array}{ccc} Gene_1^* & Gene_2 & Output \end{array} \\ \begin{bmatrix} 0 & 0 & \text{Healthy} \\ 1 & 0 & \text{Not healhy} \\ 0 & 1 & \text{Healthy} \end{bmatrix} \end{array} \tag{5.2}$$

*gene implied in the disease development.
1 - gene mutated, **0** - gene not mutated.

In order to compare the effect of mutation in a disease, a comparison between control ($Sample_0$) and the test samples ($Sample_1$ and $Sample_2$) should be done. In this example, ($Sample_0$ and $Sample_2$) are more similar, not developing the disease. Otherwise, since the samples ($Sample_0$ and $Sample_1$) have different phenotypes, they are more dissimilar.

To take into account the weight of each gene, the GDM was developed. From our data, it is considered a gene with more weight a gene with more insertions in its vicinity.

Comparing two samples, this method goes through all features to compare if, in a position i, they have the same or different value. If the value is different($mismatch$), it is multiplied by the weight of that feature. Weight ($\overrightarrow{W}$) is a vector corresponding to the number of insertions in the vicinity of a gene, across all samples. In other words, is the sum of columns of a matrix:

$$\overrightarrow{W} = \sum_{i=1}^{m} \overrightarrow{x}_i \tag{5.3}$$

where ($\overrightarrow{x_i}$) is a vector, corresponding to a sample.

The GDM distance (d_{GDM}) of two samples a and b is defined by:

$$d_{GDM}(a,b) = \sum_{i=1}^{n} \delta(a_i, b_i) \times \overrightarrow{W}_i \tag{5.4}$$

where δ is an indicator function where $\delta(a,b) = 1$ if $a \neq b$ and 0 otherwise.

Figure 5.1 shows and example of a dataset and its output using this metric.

A

		a1	a2	a3	a4	a5	a6	a7
Class1	x_1	0	0	0	0	0	0	0
	x_2	0	0	0	0	1	1	0
	x_3	0	1	0	0	0	0	0
	x_4	1	0	1	1	1	1	0
Class2	x_5	0	0	1	0	0	1	0
	x_6	0	0	1	0	1	0	0
	x_7	0	0	0	1	1	0	0
	x_8	0	0	0	0	1	0	0
Class13	x_9	0	0	0	0	0	0	0
	x_{10}	0	0	0	0	1	0	0
	x_{11}	0	0	0	0	0	0	0
	x_{12}	0	0	0	0	1	1	0
	x_{13}	0	0	0	0	0	0	0
	$\overrightarrow{W}$	1	1	3	2	7	4	0

B

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
x_1	0	11	1	17	7	10	9	7	0	7	0	11	0
x_2	11	0	12	6	10	7	6	4	11	4	11	0	11
x_3	1	12	0	18	8	11	10	8	1	8	1	12	1
x_4	17	6	18	0	10	7	8	10	17	10	17	6	17
x_5	7	10	8	10	0	11	16	14	7	14	7	10	7
x_6	10	7	11	7	11	0	5	3	10	3	10	7	10
x_7	9	6	10	8	16	5	0	2	9	2	9	6	9
x_8	7	4	8	10	14	3	2	0	7	0	7	4	7
x_9	0	11	1	17	7	10	9	7	0	7	0	11	0
x_{10}	7	4	8	10	14	3	2	0	7	0	7	4	7
x_{11}	0	11	1	17	7	10	9	7	0	7	0	11	0
x_{12}	11	0	12	6	10	7	6	4	11	4	11	0	11
x_{13}	0	11	1	17	7	10	9	7	0	7	0	11	0

Figure 5.1: Example of a dataset and its respective distance matrix using GDM metric.
A -Example of a dataset containing 13 samples, over 7 attributes, organized in 3 classes. $\overrightarrow{W}$ is a vector with the weights of a feature, in other words, the number of insertions in the vicinity of a gene. **B**- Distance matrix generated from (**A**) using the GDM distance.

The key of this metric is the weight of a gene. Let us compare the pair $d_{GDM}(x_1, x_3)$ with the pair $d_{GDM}(x_1, x_8)$. In the first pair, only one attribute is different (a_2). a_2 is a gene that was only mutated one time. On the other hand, the second pair also was one different attribute (a_5), but this gene has a large number of insertions in its vicinity. For that reason, the pair $d_{GDM}(x_1, x_8)$ has a greater distance, because the different gene seems to be really important.

5.2.2. GENE INDEPENDENT METHOD

Data contains information about integration elements in the mouse genome. Given the high dimension of the genome, only a few genes will be affected by an insertion. In fact, in the entire data, only 0.0759% of genes has insertions in its vicinity. So, when comparing two samples, it is very common to find a $match_{0-0}$ and unusual to find a $match_{1-1}$. Therefore, the $match_{1-1}$ between samples may be important.

In order to take into account the differences between all possible pairs of match and mismatch in the dataset, a new metric - the GIM - was developed. This metric calculates the probability of a $match$ and $mismatch$ happen in the whole data. This probability will be used as a weight.

Selecting two samples x_1 and x_2 the algorithm calculates the number of matches of zeros ($match_{0-0}$), matches of ones ($match_{1-1}$) and mismatches ($mismatch$):

$$\begin{aligned} match_{0-0} &= \sum_{i=1}^{n} I(x_{1i}, x_{2i}), \quad x_{1i} = 0 \\ match_{1-1} &= \sum_{i=1}^{n} I(x_{1i}, x_{2i}), \quad x_{1i} = 1 \\ mismatch &= \sum_{i=1}^{n} \delta(x_{1i}, x_{2i}) \end{aligned} \tag{5.5}$$

where n is the number of features and I and δ are indicator functions. $I(a,b) = 1$ if $a = b$ and 0 otherwise. In contrast, $\delta(a,b) = 1$ if $a \neq b$ and 0 otherwise.

The total number of matches of zeros (N_{0-0}), matches of ones (N_{1-1}) and mismatches (N_{1-0}) of the entire data is given by the sum of all the pairwises calculated.

$$\begin{aligned} N_{0-0} &= \sum_{i=1}^{p} match_{0-0\,i} \\ N_{1-1} &= \sum_{i=1}^{p} match_{1-1\,i} \\ N_{1-0} &= \sum_{i=1}^{p} mismatch_i \end{aligned} \tag{5.6}$$

where p is the number of samples' pairwises.

The total number of matches and mismatches is given by N

$$N = N_{1-1} + N_{1-0} + N_{0-0} \tag{5.7}$$

The GIM distance (d_{GIM}) of two samples a and b focus in its different elements is defined by:

$$d_{GIM}(a,b) = \frac{1}{\frac{N_{1-0}}{N}} \times \sum_{i=1}^{n} \delta(a_i, b_i) \tag{5.8}$$

A

		a1	a2	a3	a4	a5	a6	a7
Class1	x_1	0	0	0	0	0	0	0
	x_2	0	0	0	0	1	1	0
	x_3	0	1	0	0	0	0	0
	x_4	1	0	1	1	1	1	0
Class2	x_5	0	0	1	0	0	1	0
	x_6	0	0	1	0	1	0	0
	x_7	0	0	0	1	1	0	0
	x_8	0	0	0	0	1	0	0
Class13	x_9	0	0	0	0	0	0	0
	x_{10}	0	0	0	0	1	0	0
	x_{11}	0	0	0	0	0	0	0
	x_{12}	0	0	0	0	1	1	0
	x_{13}	0	0	0	0	0	0	0

B

		0 - 0	0-1/1-0	1 - 1
x_1	x_1	7	0	0
x_2	x_2	5	2	0
x_3	x_3	6	1	0
x_4	x_4	2	5	0
x_5	x_5	5	2	0
x_6	x_6	5	2	0
	⋮			
x_{13}	x_{12}	5	2	0
x_{13}	x_{13}	7	0	0

N_{x-x}	434	154	49
N	637		

C

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
x_1	0.0	8.3	4.1	20.7	8.3	8.3	8.3	4.1	0.0	4.1	0.0	8.3	0.0
x_2	8.3	0.0	12.4	12.4	8.3	8.3	8.3	4.1	8.3	4.1	8.3	0.0	8.3
x_3	4.1	12.4	0.0	24.8	12.4	12.4	12.4	8.3	4.1	8.3	4.1	12.4	4.1
x_4	20.7	12.4	24.8	0.0	12.4	12.4	12.4	16.5	20.7	16.5	20.7	12.4	20.7
x_5	8.3	8.3	12.4	12.4	0.0	8.3	16.5	12.4	8.3	12.4	8.3	8.3	8.3
x_6	8.3	8.3	12.4	12.4	8.3	0.0	8.3	4.1	8.3	4.1	8.3	8.3	8.3
x_7	8.3	8.3	12.4	12.4	16.5	8.3	0.0	4.1	8.3	4.1	8.3	8.3	8.3
x_8	4.1	4.1	8.3	16.5	12.4	4.1	4.1	0.0	4.1	0.0	4.1	4.1	4.1
x_9	0.0	8.3	4.1	20.7	8.3	8.3	8.3	4.1	0.0	4.1	0.0	8.3	0.0
x_{10}	4.1	4.1	8.3	16.5	12.4	4.1	4.1	0.0	4.1	0.0	4.1	4.1	4.1
x_{11}	0.0	8.3	4.1	20.7	8.3	8.3	8.3	4.1	0.0	4.1	0.0	8.3	0.0
x_{12}	8.3	0.0	12.4	12.4	8.3	8.3	8.3	4.1	8.3	4.1	8.3	0.0	8.3
x_{13}	0.0	8.3	4.1	20.7	8.3	8.3	8.3	4.1	0.0	4.1	0.0	8.3	0.0

Figure 5.2: Example of a dataset and its the respective distance matrix using GIM metric.
A -Example of a dataset containing 13 samples, over 7 attributes, organized in 3 classes. **B** - It counts the number of pairs existent in the dataset (N_{0-0} - 434, N_{1-0} - 154 and N_{1-1} - 49) **C**- Distance matrix generated from (**A**) using the GIM distance.

where δ is an indicator function where $\delta(a,b) = 1$ if $a \neq b$ and 0 otherwise. This transformation results in a redefine form of Hamming distance.

Figure 5.2 shows an example of a dataset and its output using GIM distance.

Example of a dataset and its respective distance matrix using GDM metric.

Let us compare the pairs $d_{GDM}(x_2, x_3)$, $d_{GDM}(x_2, x_4)$ and $d_{GDM}(x_5, x_8)$. All of them contain three mismatches and all attributes have the same weight. Once this algorithm only has in consideration the ration between mismatches values, the distance between the three pairs are the same.

The pre-processed dataset (see Section 4.3) was transformed using the following metrics: d_H, d_J, d_{GDM} and d_{GIM}. Appendix A.2 contains the heat map of the distance matrices generated by the four metrics. These transformations are used in the unsupervised an supervised learning methods.

5.3. UNSUPERVISED LEARNING

Unsupervised learning approaches reduces the size of a data, being helpful to its visualization. Feature extraction takes the original dataset and reduces its dimension preserving the proprieties of the initial data. In general, data are reduced into two dimensions (2D), so it can be visualized in a plot.

PCA [58–60] has been widely used to reduce the number of features and t-SNE [61] is a more recent approach and has shown to be very efficient. These two methods were done across the four transformation data, reducing their dimensionality into 2D. This step was done using the toolbox: Matlab Toolbox for Dimensionality Reduction[4].

5.4. SUPERVISED LEARNING

Supervised learning has been used to classify data, using a classifier. It is possible to evaluate its accuracy, using cross-validation (CV). There are dozens of different learning classifiers and each one has different performance depending the data. The ones who fits the best to the data, should be selected.

In order to select which classifiers will perform the supervised learning, two different cancer types with identical number of samples were selected. It was taken into account that both cancer types came from distinct organs. The tumour types selected were hematopoietic (140 samples) and pancreatic (228 samples).

Seven widely used classifiers were selected and tested (Table 5.1). The performance of each model was calculated using CV (k=10 and 30 repetitions). To reduce the computation time, in the next steps, the top three classifiers were selected (Figure 5.3). This top was determined by the average of the error rate of a ROC curve. The classifiers chosen for subsequent steps were NMC, kNN and SVM. This evaluation was done using the toolbox PRTools [114].

Table 5.1: Functions of PRTools used and their respective name.

Function	Classifier Name	Reference
nmc	Nearest Mean Classifier	[47]
knnc	k-Nearest Neighbour	[48]
svc	Support Vector Machine	[49]
naivebc	Naïve Bayes Classifier	[50, 51]
ldc	Linear Bayes Normal Classifier	[50]
treec	Decision Tree	[53]
randomforest	Random Forest	[54]

[4] http://lvdmaaten.github.io/drtoolbox/

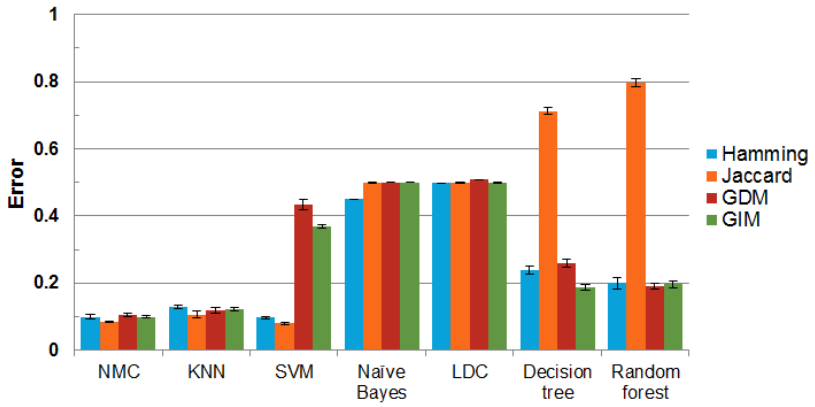


Figure 5.3: Classifiers' error rate.
Seven classifiers were tested, using two classes with identical size. The best three classifiers, with less error rate, were selected for the next steps. Mean of the error rate: NMC: 0.098; KNN: 0.119; SVM: 0.245; Naïve Bayes: 0.487; LDC: 0.502; Decision tree: 0.349; Random fores: 0.346

The supervised learning method was done across the four transformed data. Since CV evaluates the accuracy in binary classification, different subdatasets combining two tumour types were created (totalizing 21 subsets). The k-fold cross-validation was performed using the top three classifiers across those subdatasets. The number of folds applied are $k = 10$. Since the standard deviation of the first analysis has low values (Figure 5.3), the number of repetitions were reduced to fives. The output of each CV is the error rate of a ROC curve. This step was done using the toolbox PRTools [114].

5.5. GENE RANKING

The main goal of feature ranking is to sort them, according to their importance, in a given class. It is performed in two class problems. In this work, features are represented by genes.

The main aim is to find which genes are involved in a specific cancer type. For this reason, the analysis will be done comparing one class against the rest of the dataset. If samples of a tumour type has a gene containing insertions in its vicinity, and, in contrast, the rest of the samples do not have insertions for the same gene, this gene may be important for that type of cancer (Figure 5.4). This approach fits with the *diff-criterion* algorithm [73]. As an output of this algorithm, each feature will be associated to a score. Greater the score, the greater the importance of the feature to that class.

For this analysis, new subdatasets were created (totalizing 7 subdatasets). Each subdataset contains information of a tumour type and the rest cancer types are relabeled as second class. The *diff-criterion* algorithm was implemented and run through the seven subdatasets created.

Each subdataset will have it own gene scores. This score will change according to the

		a1	a2	a3	a4	a5	a6	a7
Class1	x_1	0	1	0	0	0	1	0
	x_2	0	1	1	1	1	1	0
	x_3	0	1	1	1	0	1	0
	x_4	1	1	1	1	1	1	0
Class2	x_5	0	0	1	0	0	0	1
	x_6	0	0	1	0	1	0	1
	x_7	0	0	0	1	1	1	1
	x_8	1	0	0	0	1	0	0
	$\vec{R}$	0	1	0.25	0.5	-0.25	0.75	-0.75
	Ranking Position	5	1	4	3	6	2	7

Figure 5.4: Example of feature ranking using diff-criterion algorithm.
The data given contains eight samples over seven attributes, organized in two classes. The sorted attributes according to their importance is $a_2, a_6, a_4, a_3, a_1, a_5, a_7$.

importance of the gene to a specific tumour type. Once all genes have their score, in each subdataset, the 15 greater scores are selected. This selection corresponds to the top 15 genes involved to a specific cancer type.

6

RESULTS AND DISCUSSION

The main goal of this project is to understand which genes may have an important role in the development of a specific tumour type. To achieve this goal, the data were transformed. Then, unsupervised and supervised approaches were used to find if the data have structure, allowing to distinguish different tumour types. If it is possible to differentiate different cancer types, a feature ranking may be performed, in order to find which genes may be more activated in those tumour types.

6.1. UNSUPERVISED LEARNING

Unsupervised learning methods, in particular feature extraction, allow to visualise the data into lower dimensions. Two approaches were used across the four data transformation: PCA [58–60] and t-SNE [61].

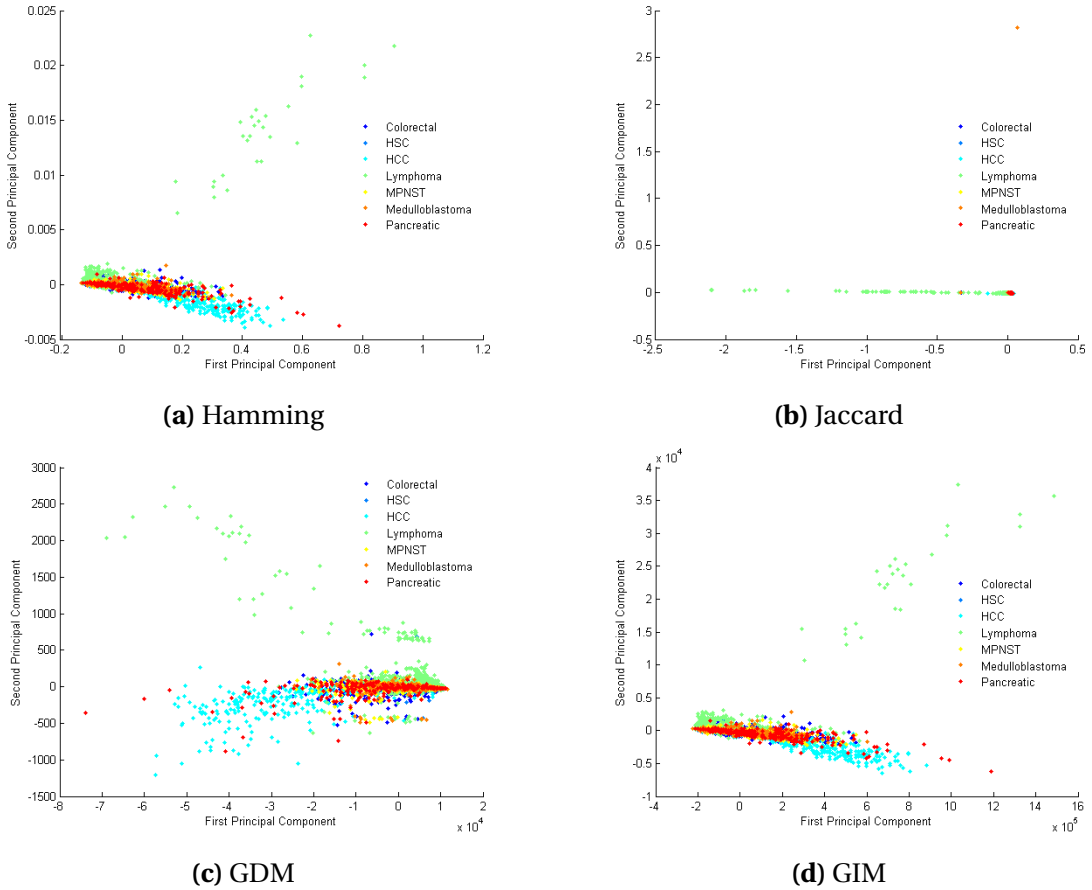
The results of PCA are shown in Figure 6.1.

The Hamming and GIM transformations have identical representations. Once both transformations have a linear correlation of 1.0, it means both represent the same. For that reason, it was expected that they had similar outputs.

The GDM seems to have a symmetric representation of GIM and Hamming transformations. In all cases, it is possible to visualise an aggregation of pancreatic tumour, as well as, HCC. The lymphoma tumour has an agglomeration, however it also has some dispersed points.

The Jaccard transformation has the most different representation. The second principal component of this transformation has low variation.

In general, PCA does not show clear clusters of the tumour types.



(a) Hamming **(b)** Jaccard

(c) GDM **(d)** GIM

Figure 6.1: Result of Principal Component Analysis (PCA) across the transformed data.
The Hamming distance and the gene independent method (GIM), seem to have identical results. On the other hand, gene dependent method (GDM) looks to be symmetric to them. The Jaccard distance shows to have the most different result. In any situation is possible not to have a clear visualization of clusters.

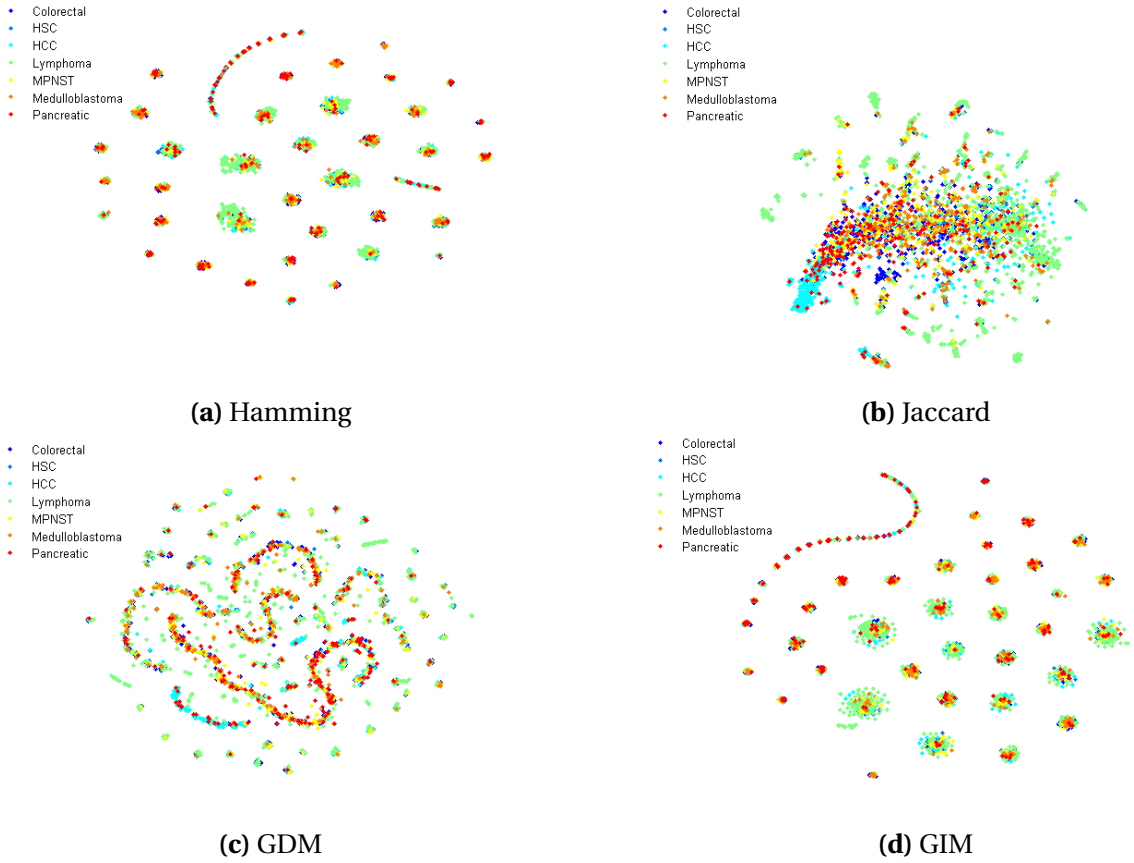
The results of t-SNE are shown in Figure 6.2.

t-SNE has a non-convex objective function. For that reason, running two times the same data will generate different representations. In fact, even GIM and Hamming transformations do not have the same illustration, they are identical. A numerous substructures are represented, prevailing the pancreatic tumour and the lymphoma tumour in the shape of some clusters.

GDM transformation creates small clusters of lymphoma tumour. However, HCC tumour seems to be agglomerate in the bottom of the figure.

Jaccard transformation has a lot of sparse points. It can happen because most of the distance between samples is equal to one. However, it is possible to visualise an aggregation of HCC tumour.

In general, t-SNE has a large overlaping tumours and it does not show clear clusters of the tumour types.



(a) Hamming **(b)** Jaccard

(c) GDM **(d)** GIM

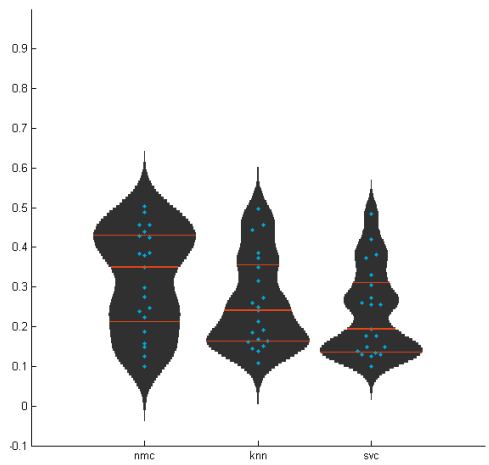
Figure 6.2: Result of t-distributed stochastic neighbor embedding (t-SNE) across the transformed data. All the transformed data show that data have some structure. However, there is an overlapping of tumour types. The Hamming distance and the gene independent method (GIM), have similar results, showing some aggregation of lymphoma and pancreatic tumour types. Gene dependent method GDM has more substructures prevailing lymphoma tumours. The Jaccard transformation has a lot of clusters of lymphoma and an HCC is clearly defined.

Both unsupervised methods do not give a lot information. PCA as a terrible visualization and t-SNE, despite showing some structure, it is not satisfied. These approaches are not sufficient. For that reason, supervised methods will be done.

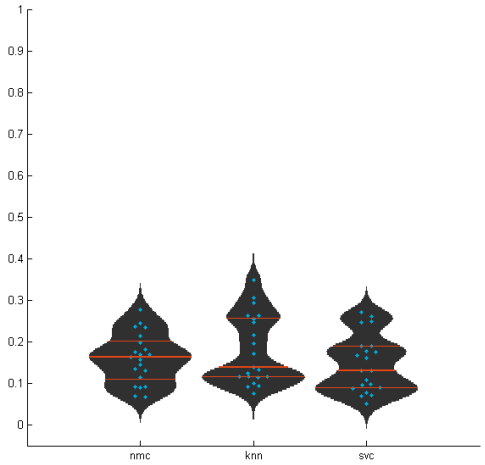
6.2. SUPERVISED LEARNING

Supervised learning methods, contrary to the unsupervised learning methods, use the samples' label for its analysis. The label corresponds to a tumour type. It was evaluated if a classifier is able to distinguish two different tumour type. Cross-validation (CV) was used to evaluate the accuracy of three classifiers - NMC, kNN and SVM -, over the four transformed data.

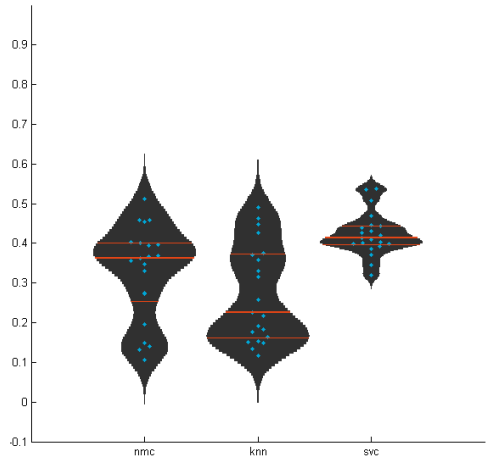
Figure 6.3 shows violins plots of the CV error rate. Each point is the value of a classifier's error rate in a binary classification. A low error rate value, is a classifier has a good accuracy. The values of the individual points in the figure are presented in Appendix B.



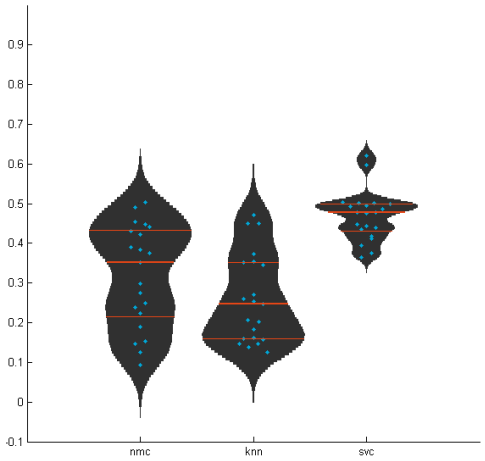
(a) Hamming - NMC: 0.32; kNN: 0.26; SVM: 0.24



(b) Jaccard - NMC: 0.16; kNN: 0.18; SVM: 0.15



(c) GDM - NMC: 0.33; kNN: 0.27; SVM: 0.42



(d) GIM - NMC: 0.32; kNN: 0.26; SVM: 0.47

Figure 6.3: Results of cross-validation (CV) across the transformated data.
Cross validation was done using three classifiers: Nearest Mean Classifier (NMC), k-Nearest Neighbour (kNN) and Support Vector Machine (SVM). The values presented are the mean of the classifier's error rate. The value of each point are listed in Apprendix-B. The black area is the probability density of the data at different values. The three red lines are the lower quartile, median and upper quartile. Jaccard transformation is the one which presents the best results.

The best classifier in Hamming transformation is SVM (0.24) and the worst is NMC. In SVM, more than 50% values are below 0.2.

The Jaccard transformation has really good results. All classifiers have an average error rate under 0.2. More than one-third of the values in SVM are under 0.1.

The best classifier in GDM and GIM is kNN (0.27 and 0.26, respectively). On the other hand, SVM shows to have the worse performance (0.42 and 0.47, respectively)

In general, classifiers have positive results over all transformations. It should be noted that SVM, in GDM and GIM transformation, has the worst results. Once GIM is a redefined way of Hamming distance, it was expected that SVM had similar performance as in NMC

and kNN. It is important to focus that Jaccard has the best performance, where the highest average error rate is 0.18. This high performance may be justified because, the Jaccard distance [70], only takes into account if genes have an insertion at least in one sample, ignoring the genes that does not have insertions in its vicinity.

Due the high dimension and complexity of the dataset, as well as, a gene can be active in several tumour types and not only a specific one, if a classifier has an error performance of 0.3 it may be an acceptable value.

6.3. RANKING GENES

Even if the unsupervised learning methods do not show a great visualization, supervised learning methods shows positive resuts. It means that the data has some relevant feature being able to distinguish differente tumour types. For that reason, it is possible to rank the genes according to their importance in a specific tumour type.

Table 6.1 shows the top 15 genes that have more insertions in its vicinity of a specific tumour type. The name of the respecitive genes are in Appendix C.

Table 6.1: List of the 15 genes more involved in a specific tumour type.
The list contains 103 genes and two of them repeat twice (Fuca1 and Ift46).

Colorectal	Hematopoietic	Hepatocellular carcinoma	Lymphoma	Malignant peripheral nerve sheat	Medulloblastoma	Pancreatic
Armc7	*Gm17535*	*Gm10337*	*Pik3r5*	*Cyp2j8*	*3110070m22Rik*	*1810049j17Rik*
Akap9	*Bc003331*	*Mrpl48*	*Gfi1*	*Pcdhga5*	*Tgif2*	*Atxn7*
Rspo2	*Ddx25*	*Rhbdd3*	*Myc*	*Pcdhgb2*	*Bcl9*	*Tmc1*
4933440n22Rik	*Pate2*	*4931422a03Rik*	*Csf2*	*Zfp106*	*Gm11273*	*Gm5129*
Fdxacb1	*Trpm8*	*Arid4a*	*Flt3*	*Ryk*	*Naip2*	*4933406p04Rik*
Baz2a	*Fam179b*	*Rtl1*	*Fuca1*	*Hdlbp*	*Prl3c1*	*Tcte2*
Samd4	*Thpo*	*Pdcd11*	*Ascl2*	*Epor*	*Akr1c20*	*Rspry1*
Prr36	*Abhd13*	*Ift46*	*Kit*	*Gpr75*	*Gm26965*	*Gm10638*
Gm10974	*Sepp1*	*Kntc1*	*Ccr7*	*Gm10542*	*Zfp87*	*Sugct*
2310061n02Rik	*Gm20537*	*Dnajb4*	*Il2rb*	*Cmya5*	*Rft1*	*Rapgefl1*
Slc6a18	*1110059e24Rik*	*Ttc4*	*Ccnd1*	*Spink11*	*Hecw1*	*4931406c07Rik*
Tshb	*Kmt2e*	*Cdkl3*	*Atp5h*	*Zfp53*	*Vmn1r200*	*Cbx5*
Alg9	*Fuca1*	*Tmem106a*	*Il3*	*Lig4*	*Gtf2h2*	*Ift46*
Matr3	*Trove2*	*Nat10*	*Snx9*	*Tbr1*	*Foxr2*	*Ccdc138*
Plod1	*Rag2*	*Cbx3*	*Eras*	*Hus1b*	*Ccl25*	*Xpnpep3*

The presented list of genes is a result of the difference between distributions of insertions in a specific class against the rest of the dataset. It selects the genes that have more insertions in its vicinity and, contrary, the rest of the dataset does not have. This approach selects the genes that are probable to be involved in a specific tumour type.

The gene list contains 85 genes already known, described in databases and annotated,

which 2 of them show up twice - *Fuca1* seems to have an important role in HSC and lymphoma tumours; and *Ift46* in HCC and pancreatic tumours. There are also 18 genes that are not annotated in the mouse genome (8 Riken complementary DNA (cDNA), 9 predicted genes and 1 cDNA sequence). Most of the listed genes were already mentioned, in the literature, as a cancer gene.

An analysis of each set of genes was done. First, to find the relationship among genes, trying to find clusters of molecular functions between them. This process was done using DAVID Database [115, 116]. A second step was know if any of the genes listed were already mentioned before in that tumour type.

The colorectal gene set has a principal function of molecular binding and involved in signalling process. Two genes have already been mentioned as likely to be involved in this type of cancer - *Akap9* [117] and *Rspo2* [118].

Genes of Hematopoietic stem cells (HSC) seem to have molecular functions related to secretion, signal and are mostly located in the membrane. *Thop* [119] and *Rag2* [120, 121] are genes suggested to affect in HSC tumour.

The Hepatocellular carcinoma (HCC) set is essentially involved in binding process. Four genes were referred to be involved in this tumour type - *Rhbdd3* [122], *Rtl1* [123], *Nat10* [124] and *Kntc* [125].

Lymphoma is the tumour type with more samples and does not have any not-annotated gene. These set of genes are participates in the leukaemia pathways and have molecular functions of binding and signalling. From the list of 15 genes, 8 were already mentioned in the literature to be included in Lymphoma's tumour- *Gfi1* [126], *Myc* [127], *Flt3* [128], *Kit* [129], *Ccr7* [130], *Il2rb* [131], *Ccnd1* [132] and *Il3* [133].

On the set of Malignant peripheral nerve sheath tumour (MPNST), no gene was mentioned before to be involved in this tumour type. However, these genes are essentially involved in binding.

Medulloblastoma genes participates in binding and transcription regulation processes. *Tgif2* [92] and *Foxr2* [89] are genes already mentioned in the literature, acting in the tumour type.

The set of pancreatic tumour are also involved in binding mechanisms. Two genes have already been mentioned as likely to be involved in this type of cancer - *Sugct* [134, 135]).

According to the data used, it is possible to generalise that each set of genes belongs to that tumour type. However, there are dozens of different types of cancer. It is quite impossible to assume that those genes are exclusive to a tumour type. But it is possible to assume that they may act on them.

More genes may be involved in tumourigenesis, being classified as oncogenes or tumour suppressor genes. Probably they are involved in several types of cancer. To find them, a multiclass analysis should be performed.

Given that at least 19 genes support the results, by published papers, these results should interest biologists. They should analyse them as new target genes, analysing both annotated and not-annotated genes.

7

CONCLUSION

7.1. OVERVIEW

The number of new cancer cases and death as a result of this disease is increasing every year. However, cancer can be prevented having a healthy lifestyle, avoid risky behaviours and get regular medical care. Do not having these prevention behaviours can create mutations in the genome. Accumulating those mutations a cell can develop the process to lead to cancer. The data used in this work are a collection of independent studies of IM in mice. These integrations can increase or decrease the gene expression on its location.

Machine learning has been a popular strategy for medical researchers. Different approaches can be able to discover and identify patterns in complex datasets. In this work, different techniques were used. Four dissimilarity approaches, two of them were developed to take into account the biological meaning of the data. Two unsupervised learning method and supervised learning were used. Once these approaches validate the structure of the data, the feature ranking may be performed.

The main results are set list of the 15 genes for each tumour type. Each one contains the genes that are probable to be more involved in that tumour type. In summary, 103 genes were listed, where 18 of them are not annotated. After a bibliographic research, at least 19 genes were already mentioned in the literature to act in that tumour type. For that reason, these sets of genes can, apparently, be involved in the tumourigenesis process.

7.2. LIMITATIONS

As all scientific work has its limitations, no exceptions in this case were observed. The main limitations of this work are described below:

- **Transformation data -** The gene independent method (GIM) should takes into account the distribution of $match_{0-0}$ and $match_{1-1}$, giving more weight to the $match_{1-1}$. In fact, this does not happen with the algorithm, creating, therefore, a re-formulation of the Hamming distance.
- **Unsupervised learning -** both PCA and t-SNE visualization are unclear. Testing other unsupervised learning methods could be helpful.
- **Supervised learning -** SVM, NMC and kNN have good performance when used in Jaccard transformation. Nonetheless, in the other transformations, these classifiers have a bit more error rate. It would be interesting to test more models, or even create a new one more adapted to the structure of the data.

7.3. RECOMMENDATIONS

Although the main goals proposed for this project have been accomplished, some features could be added to improve the results:

- Increase the window. It is known that integrations can affect genes in long distances [35]. Increasing the size of windows, the window space will capture more insertions, having more informative data and avoiding samples to have no insertions in its vicinity;
- Take into account the position of an insertion. In fact, the integration can be in the vicinity of a gene, or within it having contradictory results. If an insertion is the vicinity of a gene, it will be more active. On the other hand, if an insertion is within the gene, it will deregulate it. Knowing their integration position will help to distinguish between oncogenes and tumour suppressor genes;
- Validation of the results *in vitro* and *in vivo*. Validating genes that are involved in tumourigenesis, researchers can use drug therapy to correct abnormal gene activity and prevent the development of cancer.

BIBLIOGRAPHY

[1] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, *Global cancer statistics,* CA: A Cancer Journal for Clinicians **61**, 69 (2011).

[2] L. a. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-tieulent, and A. Jemal, *Global cancer statistics, 2012,* CA: A Cancer Journal for Clinicians **65**, 87 (2015).

[3] T.-M. Huang, V. Kecman, and I. Kopriva, *Kernel based algorithms for mining huge data sets*, 1st ed., Vol. 17 (Springer, 2006) p. 260.

[4] Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. X. Mayer, and H. W. Mewes, *Gene selection from microarray data for cancer classification - A machine learning approach,* Computational Biology and Chemistry **29**, 37 (2005).

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Gene gelection for cancer classification using Support Vector Machines,* Mach. Learn. **46**, 389 (2002).

[6] A. C. Tan and D. Gilbert, *Ensemble Machine Learnign on gene expression data for cancer classification,* Applied bioinformatics **2**, 1 (2003).

[7] E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, and S. Canaider, *An estimation of the number of cells in the human body.* Annals of human biology **40**, 463 (2013).

[8] J. C. Roach, G. Glusman, A. F. a. Smit, C. D. Huff, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and J. Galas, *Analysis of genetic inheritance in a family quartet by whole genome sequencing,* NIH Public Access **328**, 636 (2010).

[9] R. Doll and R. Peto, *The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today,* Journal of the National Cancer Institute **66**, 1191 (1981).

[10] K. Czene, P. Lichtenstein, and K. Hemminki, *Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database,* International journal of cancer. **99**, 260 (2002).

[11] W. P. Roos and B. Kaina, *DNA damage-induced cell death by apoptosis,* Trends in Molecular Medicine **12**, 440 (2006).

[12] J. A. Karam and J.-T. Hsieh, *Anti-cancer strategy of transitional cell carcinoma of bladder based on induction of different types of programmed cell deaths,* in *Apoptosis in Carcinogenesis and Chemotherapy*, Vol. 1 (Springer, 2009) pp. 25–50.

[13] L. H. Hartwell and T. a. Weinert, *Checkpoints: controls that ensure the order of cell cycle events,* Science New York **246**, 629 (1989).

[14] P. C. Nowell, *The clonal evolution of tumor cell populations,* Science **194**, 23 (1976).

[15] M. Greaves, *Cancer causation: The Darwinian downside of past success?* Lancet Oncology **3**, 244 (2002).

[16] M. R. Stratton, P. J. Campbell, and P. A. Futreal, *The cancer genome.* Nature **458**, 719 (2009).

[17] I. Tomlinson, M. Novelli, and W. Bodmer, *The mutation rate and cancer,* Proceedings of the National Academy of Sciences **93**, 14800 (1996).

[18] L. A. Loeb, *A mutator phenotype in cancer,* Cancer research **61**, 3230 (2001).

[19] D. Hanahan, R. a. Weinberg, and S. Francisco, *The hallmarks of cancer,* Cell **100**, 57 (2000).

[20] D. Hanahan and R. a. Weinberg, *Hallmarks of cancer: The next generation,* Cell **144**, 646 (2011).

[21] G. P. Gupta and J. Massagué, *Cancer Metastasis: Building a Framework,* Cell **127**, 679 (2006).

[22] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, *A census of human cancer genes.* Nature reviews. Cancer **4**, 177 (2004).

[23] J. D. Rowley, *Biological implications of consistent chromosome rearragements in leukemia and lymphoma,* Cancer Research **44**, 3159 (1984).

[24] A. G. Knudson, *Mutation and cancer: statistical study of retinoblastoma.* Proceedings of the National Academy of Sciences of the United States of America **68**, 820 (1971).

[25] R. Sager, *Genetic suppression of tumor formation: a new frontier in cancer research,* Cancer Research **46**, 1573 (1986).

[26] J. Mattison, L. van der Weyden, T. Hubbard, and D. J. Adams, *Cancer gene discovery in mouse and man.* Biochimica et biophysica acta **1796**, 140 (2009).

[27] R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigo, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, and A. Stabenau, *Initial sequencing and comparative analysis of the mouse genome,* Nature **420**, 520 (2002).

[28] S. J. Howe, M. R. Mansour, K. Schwarzwaelder, M. Hubank, H. Kempski, M. H. Brugman, D. D. Ridder, K. C. Gilmour, S. Adams, S. I. Thornhill, K. L. Parsley, F. J. T. Staal, E. Rosemary, D. C. Linch, J. Bayford, L. Brown, M. Quaye, C. Kinnon, P. Ancliff, D. K. Webb, M. Schmidt, V. Kalle, H. B. Gaspar, and A. J. Thrasher, *Insertional mutagenesis in combination with acquired somatic mutations leads to leukemogenesis following gene therapy of SCID-X1,* The Journal of Clinical Investigation **44**, 3143 (2008).

[29] J. M. Alonso, A. N. Stepanova, T. J. Leisse, C. J. Kim, H. Chen, P. Shinn, D. K. Stevenson, J. Zimmerman, P. Barajas, R. Cheuk, C. Gadrinab, C. Heller, A. Jeske, E. Koesema, C. C. Meyers, H. Parker, L. Prednis, Y. Ansari, N. Choy, H. Deen, M. Geralt, N. Hazari, E. Hom, M. Karnes, C. Mulholland, R. Ndubaku, I. Schmidt, P. Guzman, L. Aguilar-Henonin, M. Schmid, D. Weigel, D. E. Carter, T. Marchand, E. Risseeuw, D. Brogden, A. Zeko, W. L. Crosby, C. C. Berry, and J. R. Ecker, *Genome-wide insertional mutagenesis of Arabidopsis thaliana.* Science (New York, N.Y.) **301**, 653 (2003).

[30] V. Pečenka, P. Pajer, V. Karafiát, and M. Dvořák, *Chicken models of retroviral insertional mutagenesis,* in *Insertional Mutagenesis Strategies in Cancer Genetics*, Vol. 1 (Springer, 2011) pp. 77–112.

[31] A. Amsterdam, S. Burgess, G. Golling, W. Chen, Z. Sun, K. Townsend, S. Farrington, M. Haldi, and N. Hopkins, *A large-scale insertional mutagenesis screen in zebrafish,* Genes and Development **13**, 2713 (1999).

[32] L. Cooley, R. Kelley, and A. Spradling, *Insertional Mutagenesis of the Drosophila Genome with Single P Elements,* Science **239**, 1121 (1988).

[33] J. S. Jeon, S. Lee, K. H. Jung, S. H. Jun, D. H. Jeong, J. Lee, C. Kim, S. Jang, S. Lee, K. Yang, J. Nam, K. An, M. J. Han, R. J. Sung, H. S. Choi, J. H. Yu, J. H. Choi, S. Y. Cho, S. S. Cha, S. I. Kim, and G. An, *T-DNA insertional mutagenesis for functional genomics in rice,* Plant Journal **22**, 561 (2000).

[34] E. D. Mullins, X. Chen, P. Romaine, R. Raina, D. M. Geiser, and S. Kang, *Agrobacterium-mediated transformation of Fusarium oxysporum: an efficient tool for insertional mutagenesis and gene transfer.* Phytopathology **91**, 173 (2001).

[35] P. A. Lazo, J. S. Lee, and P. N. Tsichlis, *Long-distance activation of the Myc protooncogene by provirus insertion in Mlvi-1 or Mlvi-4 in rat T-cell lymphomas.* Proceedings of the National Academy of Sciences of the United States of America **87**, 170 (1990).

[36] A. Berns, *Insertional Mutagenesis: a powerful tool in cancer research,* in *Insertional Mutagenesis strategies in cancer genetics* (Springer, New York, United States, 2011) pp. 1–18.

[37] J. Kool and A. Berns, *High-throughput insertional mutagenesis screens in mice to identify oncogenic networks.* Nature reviews. Cancer **9**, 389 (2009).

[38] J. de Jong, W. Akhtar, J. Badhai, A. G. Rust, R. Rad, J. Hilkens, A. Berns, M. van Lohuizen, L. F. a. Wessels, and J. de Ridder, *Chromatin Landscapes of Retroviral and Transposon Integration Profiles,* PLoS Genetics **10**, 17 (2014).

[39] P. Larranaga, *Machine learning in bioinformatics,* Briefings in Bioinformatics **7**, 86 (2006).

[40] I. Kononenko, *Machine learning for medical diagnosis: history, state of the art and perspective,* Artificial Intelligence in Medicine **23**, 89 (2001).

[41] A. Khotanzad and Y. H. Hong, *Invariant image recognition by Zernike moments,* IEEE Transactions on Pattern Analysis and Machine Intelligence **12**, 489 (1990).

[42] L. Deng and X. Li, *Machine learning paradigms for speech recognition: An overview,* IEEE Transactions on Audio, Speech and Language Processing **21**, 1 (2013).

[43] F. Sebastiani, *Machine Learning in Automated Text Categorization,* ACM computing surveys (CSUR) **34**, 1 (2002).

[44] S. Dubnov, G. Assayag, O. Lartillot, and G. Bejerano, *Using Machine-Learning methods for musical style modeling,* Computer **36**, 3 (2003).

[45] A. L. Samuel, *Some studies in machine learning using the game of checkers,* IBM Journal of research and development **3**, 210 (1959).

[46] M. Mohri and A. Rostamizadeh, Afshin; Talwalkar, *Foundations of machine learning* (MIT press, 2012).

[47] P. Datta and D. F. Kibler, *Symbolic Nearest Mean Classifiers,* Association for the Advancement of Artificial Intelligence , 82 (1997).

[48] T. Cover and P. Hart, *Nearest neighbor pattern classification,* IEEE Transactions on Information Theory **13**, 21 (1967).

[49] C. Cortes and V. Vapnik, *Support-Vector Networks,* Machine Learning **20**, 273 (1995).

[50] R. O. Duda, P. E. Hart, *et al.*, *Pattern classification and scene analysis*, Vol. 3 (Wiley New York, 1973).

[51] P. Langley, W. Iba, and K. Thompson, *An Analysis of Bayesian Classiers,* Association for the Advancement of Artificial Intelligence **90**, 1 (1992).

[52] T. M. Mitchell, *Machine Learning*, 1st ed., 1 (McGraw Hill, Portland, 1997) pp. 154–200.

[53] J. R. Quinlan, *Induction of decision trees,* Machine Learning **1**, 81 (1986).

[54] L. Breiman, *Random forests,* Machine learning **45**, 5 (2001).

[55] N. R. Draper and H. Smith, *Applied regression analysis*, 3rd ed. (John Wiley & Sons, Inc., Hoboken, NJ, USA, 1998).

[56] A. K. Jain, M. N. Murty, and P. J. Flynn, *Data clustering: a review,* ACM Computing Surveys **31**, 264 (1999).

[57] V. Estivill-Castro, *Why so many clustering algorithms: a position paper,* ACM SIGKDD Explorations Newsletter **4**, 65 (2002).

[58] K. Pearson, *On lines and planes of closest fit to systems of points in space,* The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**, 559 (1901).

[59] H. Hotelling, *Analysis of a complex of statistical variables into Principal Components,* Journal of Educational Psychology **5**, 417 (1933).

[60] I. T. Jolliffe, *Encyclopedia of Statistics in behavioral science*, 2nd ed., Vol. 30 (Springer Series in Statistics, New York, 2002).

[61] L. Van Der Maaten and G. Hinton, *Visualizing Data using t-SNE,* Journal of Machine Learning Research **9**, 2579 (2008).

[62] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition,* Proceedings of the IEEE **86**, 2278 (1998).

[63] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed., Vol. 1 (Springer-Verlag New York, 2009).

[64] R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection,* International Joint Conference on Artificial Intelligence **14**, 1 (1995).

[65] M. H. Zweig and G. Campbell, *Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine,* Clinical Chemistry **39**, 561 (1993).

[66] A. R. Webb, *Measures of dissimilarity,* in *Statistical pattern recognition* (John Wiley & Sons, Malvern, 2003) 2nd ed., pp. 419–429.

[67] W. M. Fitch and E. Margoliash, *Construction of phylogenetic trees,* Science **155**, 279 (1967).

[68] L. L. Cavalli-Sforza and a. W. F. Edwards, *Phylogenetic analysis. Models and estimation procedures,* The American Journal of Human Genetics **19**, 233 (1967).

[69] R. Hamming, *Error detecting and error correcting codes,* Bell System Technical Journal **29**, 147 (1950).

[70] P. Jaccard, *The distribution of the flora in the alphine zone,* The New Phytologist **11**, 37 (1912).

[71] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*, 1st ed. (Springer-Verlag Berlin Heidelberg, 2006).

[72] K. Javed, M. Saeed, and H. a. Babri, *The correctness problem: evaluating the ordering of binary features in rankings,* Knowledge and Information Systems **39**, 543 (2013).

[73] K. Javed, H. A. Babri, and M. Saeed, *Feature selection based on class-dependent densities for high-dimensional binary data,* IEEE Transactions on Knowledge and Data Engineering **24**, 465 (2012).

[74] D. P. Ryan, T. S. Hong, and N. Bardeesy, *Pancreatic adenocarcinoma,* New England Journal of Medicine **371**, 1039 (2014).

[75] E. A. Bard-Chapeau, A.-T. Nguyen, A. G. Rust, A. Sayadi, P. Lee, B. Q. Chua, L.-S. New, J. de Jong, J. M. Ward, C. K. Y. Chin, V. Chew, H. C. Toh, J.-P. Abastado, T. Benoukraf, R. Soong, F. a. Bard, A. J. Dupuy, R. L. Johnson, G. K. Radda, E. C. Y. Chan, L. F. a. Wessels, D. J. Adams, N. a. Jenkins, and N. G. Copeland, *Transposon mutagenesis identifies genes driving hepatocellular carcinoma in a chronic hepatitis B mouse model.* Nature genetics **46**, 24 (2014).

[76] A. M. Bender, L. S. Collier, F. J. Rodriguez, C. Tieu, J. D. Larson, C. Halder, E. Mahlum, T. M. Kollmeyer, K. Akagi, G. Sarkar, D. A. Largaespada, and R. B. Jenkins, *Sleeping beauty-mediated somatic mutagenesis implicates CSF1 in the formation of high-grade astrocytomas,* Cancer Research **70**, 3557 (2010).

[77] K. E. Berquam-Vrieze, K. Nannapaneni, B. T. Brett, L. Holmfeldt, M. Jing, O. Zagorodna, N. A. Jenkins, N. G. Copeland, D. K. Meyerholz, C. Michael Knudson, C. G. Mullighan, T. E. Scheetz, and A. J. Dupuy, *Cell of origin strongly influences genetic selection in a mouse model of T-ALL,* Blood **118**, 4646 (2011).

[78] D. Cesana, M. Ranzani, M. Volpin, C. Bartholomae, C. Duros, A. Artus, S. Merella, F. Benedicenti, L. Sergi Sergi, F. Sanvito, C. Brombin, A. Nonis, C. D. Serio, C. Doglioni, C. von Kalle, M. Schmidt, O. Cohen-Haguenauer, L. Naldini, and E. Montini, *Uncovering and dissecting the genotoxicity of self-inactivating lentiviral vectors in vivo.* Molecular therapy **22**, 774 (2014).

[79] L. S. Collier, D. J. Adams, C. S. Hackett, L. E. Bendzick, K. Akagi, M. N. Davies, M. D. Diers, F. J. Rodriguez, A. M. Bender, C. Tieu, I. Matise, A. J. Dupuy, N. G. Copeland, N. a. Jenkins, J. G. Hodgson, W. a. Weiss, R. B. Jenkins, and D. a. Largaespada, *Whole-body sleeping beauty mutagenesis can cause penetrant leukemia/lymphoma and rare high-grade glioma without associated embryonic lethality,* Cancer Research **69**, 8429 (2009).

[80] L. S. Collier, C. M. Carlson, S. Ravimohan, A. J. Dupuy, and D. A. Largaespada, *Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse.* Nature **436**, 272 (2005).

[81] A. J. Dupuy, L. M. Rogers, J. Kim, K. Nannapaneni, T. K. Starr, P. Liu, D. A. Largaespada, T. E. Scheetz, N. A. Jenkins, and N. G. Copeland, *A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice,* Cancer Research **69**, 8150 (2009).

[82] A. J. Dupuy, K. Akagi, D. a. Largaespada, N. G. Copeland, and N. a. Jenkins, *Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system.* Nature **436**, 221 (2005).

[83] R. H. Friedel, C. C. Friedel, T. Bonfert, R. Shi, R. Rad, and P. Soriano, *Clonal expansion analysis of transposon insertions by high-throughput sequencing identifies candidate cancer genes in a Piggybac mutagenesis screen,* PLOS ONE **8**, 1 (2013).

[84] L. A. Genovesi, C. G. Ng, M. J. Davis, M. Remke, M. D. Taylor, D. J. Adams, A. G. Rust, J. M. Ward, K. H. Ban, N. A. Jenkins, N. G. Copeland, and B. J. Wainwright, *Sleeping Beauty mutagenesis in a mouse medulloblastoma model defines networks that discriminate between human molecular subgroups.* Proceedings of the National Academy of Sciences of the United States of America **110**, E4325 (2013).

[85] C. A. Huser, K. L. Gilroy, J. de Ridder, A. Kilbey, G. Borland, N. Mackay, A. Jenkins, M. Bell, P. Herzyk, L. van der Weyden, D. J. Adams, A. G. Rust, E. Cameron, and J. C. Neil, *Insertional mutagenesis and deep profiling reveals gene hierarchies and a Myc/p53-dependent bottleneck in lymphomagenesis,* PLoS Genetics **10** (2014).

[86] V. W. Keng, D. Sia, A. L. Sarver, B. R. Tschida, D. Fan, C. Alsinet, M. Solé, W. L. Lee, T. P. Kuka, B. S. Moriarity, A. Villanueva, A. J. Dupuy, J. D. Riordan, J. B. Bell, K. a. Kevin, J. M. Llovet, and D. a. Largaespada, *Sex bias occurrence of hepatocellular carcinoma in Poly7 molecular subclass is associated with EGFR,* Hepatology **57**, 120 (2013).

[87] V. W. Keng, A. Villanueva, D. Y. Chiang, A. J. Dupuy, B. J. Ryan, I. Matise, K. a. T. Silverstein, A. Sarver, T. K. Starr, K. Akagi, L. Tessarollo, L. S. Collier, S. Powers, S. W. Lowe, N. a. Jenkins, N. G. Copeland, J. M. Llovet, and D. a. Largaespada, *A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma.* Nature biotechnology **27**, 264 (2009).

[88] J. Kool, A. G. Uren, C. P. Martins, D. Sie, J. de Ridder, G. Turner, M. van Uitert, K. Matentzoglu, W. Lagcher, P. Krimpenfort, J. Gadiot, C. Pritchard, J. Lenz, A. H. Lund, J. Jonkers, J. Rogers, D. J. Adams, L. Wessels, A. Berns, and M. van Lohuizen, *Insertional mutagenesis in mice deficient for p15Ink4b, p16 Ink4a, p21Cip1, and p27Kip1 reveals cancer gene interactions and correlations with tumor phenotypes,* Cancer Research **70**, 530 (2010).

[89] H. Koso, A. Tsuhako, E. Lyons, J. M. Ward, A. G. Rust, D. J. Adams, N. A. Jenkins, N. G. Copeland, and S. Watanabe, *Identification of FoxR2 as an oncogene in medulloblastoma,* Cancer Research **74**, 2351 (2014).

[90] H. Koso, H. Takeda, C. Chin, K. Yew, J. M. Ward, N. Nariai, and K. Ueno, *Transposon mutagenesis identi fi es genes that transform neural stem cells into glioma-initiating cells,* Proceedings of the National Academy of Sciences **109** (2012).

[91] M. J. Koudijs, C. Klijn, L. van der Weyden, J. Kool, J. Ten Hoeve, D. Sie, P. R. Prasetyanti, E. Schut, S. Kas, T. Whipp, E. Cuppen, L. Wessels, D. J. Adams, and J. Jonkers, *High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors,* Genome Research **21**, 2181 (2011).

[92] M. Lastowska, H. Al-Afghani, H. H. Al-Balool, H. Sheth, E. Mercer, J. M. Coxhead, C. P. Redfern, H. Peters, A. D. Burt, M. Santibanez-Koref, C. M. Bacon, L. Chesler, A. G. Rust, D. J. Adams, D. Williamson, S. C. Clifford, and M. S. Jackson, *Identification of*

a neuronal transcription factor network involved in medulloblastoma development. Acta neuropathologica communications **1**, 35 (2013).

[93] K. M. Mann, J. M. Ward, C. C. K. Yew, A. Kovochich, D. W. Dawson, M. A. Black, B. T. Brett, T. E. Sheetz, A. J. Dupuy, D. K. Chang, A. V. Biankin, N. Waddell, K. S. Kassahn, S. M. Grimmond, A. G. Rust, D. J. Adams, N. A. Jenkins, and N. G. Copeland, *Inaugural Article: Sleeping Beauty mutagenesis reveals cooperating mutations and pathways in pancreatic adenocarcinoma,* Proceedings of the National Academy of Sciences **109**, 5934 (2012).

[94] H. N. March, A. G. Rust, N. A. Wright, J. ten Hoeve, J. de Ridder, M. Eldridge, L. van der Weyden, A. Berns, J. Gadiot, A. Uren, R. Kemp, M. J. Arends, L. F. a. Wessels, D. J. Winton, and D. J. Adams, *Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis,* Nature Genetics **43**, 1202 (2011).

[95] K. A. O'Donnell, V. W. Keng, B. York, E. L. Reineke, D. Seo, D. Fan, K. A. T. Silverstein, C. T. Schrum, W. R. Xie, L. Mularoni, S. J. Wheelan, M. S. Torbenson, B. W. O'Malley, D. a. Largaespada, and J. D. Boeke, *A Sleeping Beauty mutagenesis screen reveals a tumor suppressor role for Ncoa2/Src-2 in liver cancer,* Proceedings of the National Academy of Sciences **109**, E1377 (2012).

[96] P. A. Pérez-Mancera, A. G. Rust, L. van der Weyden, G. Kristiansen, A. Li, A. L. Sarver, K. A. T. Silverstein, R. Grützmann, D. Aust, P. Rümmele, T. Knösel, C. Herd, D. L. Stemple, R. Kettleborough, J. A. Brosnan, A. Li, R. Morgan, S. Knight, J. Yu, S. Stegeman, L. S. Collier, J. J. ten Hoeve, J. de Ridder, A. P. Klein, M. Goggins, R. H. Hruban, D. K. Chang, A. V. Biankin, S. M. Grimmond, A. V. Biankin, A. L. Johns, A. Mawson, D. K. Chang, M.-A. L. Brancato, S. J. Rowe, S. L. Simpson, M. Martyn-Smith, L. a. Chantrill, V. T. Chin, A. Chou, M. J. Cowley, J. L. Humphris, M. D. Jones, R. Scott Mead, A. M. Nagrial, M. Pajic, J. Pettit, M. Pinese, I. Rooman, J. Wu, R. J. Daly, E. A. Musgrove, R. L. Sutherland, S. M. Grimmond, N. Waddell, K. S. Kassahn, D. K. Miller, P. J. Wilson, A.-M. Patch, S. Song, I. Harliwong, S. Idrisoglu, C. Nourse, E. Nourbakhsh, S. Manning, S. Wani, M. Gongora, M. Anderson, O. Holmes, C. Leonard, D. Taylor, S. Wood, C. Xu, K. Nones, J. Lynn Fink, A. Christ, T. Bruxner, N. Cloonan, F. Newell, J. V. Pearson, J. S. Samra, A. J. Gill, N. Pavlakis, A. Guminski, C. Toon, A. V. Blankin, R. Asghari, N. D. Merrett, D. K. Chang, D. a. Pavey, A. Das, P. H. Cosman, K. Ismail, C. O'Connor, V. W. Lam, D. McLeod, H. C. Pleass, V. James, J. G. Kench, C. L. Cooper, D. Joseph, C. Sandroussi, M. Crawford, M. Texler, C. Forrest, A. Laycock, K. P. Epari, M. Ballal, D. R. Fletcher, S. Mukhedkar, N. a. Spry, B. DeBoer, M. Chai, K. Feeney, N. Zeps, M. Beilin,

N. Q. Nguyen, A. R. Ruszkiewicz, C. Worthley, C. P. Tan, T. Debrencini, J. Chen, M. E. Brooke-Smith, V. Papangelis, H. Tang, A. P. Barbour, A. D. Clouston, P. Martin, T. J. O'Rourke, A. Chiang, J. W. Fawcett, K. Slater, S. Yeung, M. Hatzifotis, P. Hodgkinson, C. Christophi, M. Nikfarjam, V. Cancer Biobank, J. R. Eshleman, R. H. Hruban, A. Maitra, C. a. Iacobuzio-Donahue, R. D. Schulick, C. L. Wolfgang, R. a. Morgan, R. T. Lawlor, S. Beghelli, V. Corbo, M. Scardoni, C. Bassi, M. a. Tempero, L. F. a. Wessels, S. a. Wood, C. a. Iacobuzio-Donahue, C. Pilarsky, D. a. Largaespada, D. J. Adams, and D. a. Tuveson, *The deubiquitinase USP9X suppresses pancreatic ductal adenocarcinoma,* Nature **486**, 266 (2012).

[97] R. M. Quintana, A. J. Dupuy, A. Bravo, M. L. Casanova, J. P. Alameda, A. Page, M. Sánchez-Viera, A. Ramírez, and M. Navarro, *A transposon-based analysis of gene mutations related to skin cancer development.* The Journal of investigative dermatology **133**, 239 (2013).

[98] R. Rad, L. Rad, W. Wang, J. Cadinanos, G. Vassiliou, S. Rice, L. S. Campos, K. Yusa, R. Banerjee, M. A. Li, J. de la Rosa, A. Strong, D. Lu, P. Ellis, N. Conte, F. T. Yang, P. Liu, and A. Bradley, *PiggyBac transposon mutagenesis: a tool for cancer gene discovery in mice.* Science (New York, N.Y.) **330**, 1104 (2010).

[99] E. P. Rahrmann, A. L. Watson, V. W. Keng, K. Choi, B. S. Moriarity, D. A. Beckmann, N. K. Wolf, A. Sarver, M. H. Collins, C. L. Moertel, M. R. Wallace, B. Gel, E. Serra, N. Ratner, and D. A. Largaespada, *Forward genetic screen for malignant peripheral nerve sheath tumor formation identifies new genes and pathways driving tumorigenesis.* Nature Genetics **45**, 756 (2013).

[100] M. Ranzani, D. Cesana, C. C. Bartholomae, F. Sanvito, M. Pala, F. Benedicenti, P. Gallina, L. S. Sergi, S. Merella, A. Bulfone, C. Doglioni, C. von Kalle, Y. J. Kim, M. Schmidt, G. Tonon, L. Naldini, and E. Montini, *Lentiviral vector-based insertional mutagenesis identifies genes associated with liver cancer.* Nature methods **10**, 155 (2013).

[101] T. K. Starr, P. M. Scott, B. M. Marsh, L. Zhao, B. L. N. Than, M. G. O'Sullivan, A. L. Sarver, A. J. Dupuy, D. A. Largaespada, and R. T. Cormier, *A Sleeping Beauty transposon-mediated screen identifies murine susceptibility genes for adenomatous polyposis coli (Apc)-dependent intestinal tumorigenesis.* Proceedings of the National Academy of Sciences of the United States of America **108**, 5765 (2011).

[102] T. K. Starr, R. Allaei, K. A. T. Silverstein, R. A. Staggs, A. L. Sarver, T. L. Bergemann, M. Gupta, M. G. O'Sullivan, I. Matise, A. J. Dupuy, L. S. Collier, S. Powers, A. L. Oberg,

Y. W. Asmann, S. N. Thibodeau, L. Tessarollo, N. G. Copeland, N. A. Jenkins, R. T. Cormier, and D. A. Largaespada, *A transposon-based genetic screen in mice identifies genes altered in colorectal cancer.* Science (New York, N.Y.) **323**, 1747 (2009).

[103] V. Theodorou, M. A. Kimm, M. Boer, L. Wessels, W. Theelen, J. Jonkers, and J. Hilkens, *MMTV insertional mutagenesis identifies genes, gene families and pathways involved in mammary cancer.* Nature genetics **39**, 759 (2007).

[104] A. G. Uren, J. Kool, K. Matentzoglu, J. de Ridder, J. Mattison, M. van Uitert, W. Lagcher, D. Sie, E. Tanger, T. Cox, M. Reinders, T. J. Hubbard, J. Rogers, J. Jonkers, L. Wessels, D. J. Adams, M. van Lohuizen, and A. Berns, *Large-scale mutagenesis in p19ARF- and p53-deficient mice identifies cancer genes and their collaborative Nnetworks,* Cell **133**, 727 (2008).

[105] L. van der Weyden, G. Giotopoulos, A. G. Rust, L. S. Matheson, F. W. V. Delft, J. Kong, A. E. Corcoran, M. F. Greaves, C. G. Mullighan, B. J. Huntly, and D. J. Adams, *Modeling the evolution of ETV6-RUNX1 - induced B-cell precursor acute lymphoblastic leukemia in mice,* Blood **118**, 1041 (2011).

[106] L. van der Weyden, A. Papaspyropoulos, G. Poulogiannis, A. G. Rust, M. Rashid, D. J. Adams, M. J. Arends, and E. O'Neill, *Loss of Rassf1a synergizes with deregulated Runx2 signaling in tumorigenesis,* Cancer Research **72**, 3817 (2012).

[107] L. van der Weyden, M. J. Arends, A. G. Rust, G. Poulogiannis, R. E. McIntyre, and D. J. Adams, *Increased tumorigenesis associated with loss of the tumor suppressor gene Cadm1,* Molecular Cancer **11**, 29 (2012).

[108] L. van der Weyden, A. G. Rust, R. E. McIntyre, C. D. Robles-Espinoza, M. del Castillo Velasco-Herrera, R. Strogantsev, A. C. Ferguson-Smith, S. McCarthy, T. M. Keane, M. J. Arends, and D. J. Adams, *Jdp2 downregulates Trp53 transcription to promote leukaemogenesis in the context of Trp53 heterozygosity,* Oncogene **32**, 397 (2012).

[109] G. S. Vassiliou, J. L. Cooper, R. Rad, J. Li, S. Rice, A. Uren, L. Rad, P. Ellis, R. Andrews, R. Banerjee, C. Grove, W. Wang, P. Liu, P. Wright, M. Arends, and A. Bradley, *Mutant nucleophosmin and cooperating pathways drive leukemia initiation and progression in mice.* Nature genetics **43**, 470 (2011).

[110] C. C. Wong, I. Martincorena, A. G. Rust, M. Rashid, C. Alifrangis, L. B. Alexandrov, J. C. Tiffen, C. Kober, A. R. Green, C. E. Massie, J. Nangalia, S. Lempidaki, H. Döhner, K. Döhner, S. J. Bray, U. McDermott, E. Papaemmanuil, P. J. Campbell, and D. J.

Adams, *Inactivating CUX1 mutations promote tumorigenesis.* Nature genetics **46**, 33 (2014).

[111] X. Wu, P. A. Northcott, A. Dubuc, A. J. Dupuy, D. J. H. Shih, H. Witt, S. Croul, E. Bouffet, D. W. Fults, C. G. Eberhart, L. Garzia, T. Van Meter, D. Zagzag, N. Jabado, J. Schwartzentruber, J. Majewski, T. E. Scheetz, S. M. Pfister, A. Korshunov, X.-N. Li, S. W. Scherer, Y.-J. Cho, K. Akagi, T. J. MacDonald, J. Koster, M. G. McCabe, A. L. Sarver, V. P. Collins, W. A. Weiss, D. A. Largaespada, L. S. Collier, and M. D. Taylor, *Clonal selection drives genetic divergence of metastatic medulloblastoma,* Nature **482**, 529 (2012).

[112] N. Zanesi, V. Balatti, J. Riordan, A. Burch, L. Rizzotto, A. Palamarchuk, L. Cascione, A. Lagana, A. J. Dupuy, C. M. Croce, Y. Pekarsky, and W. Dc, *A Sleeping Beauty screen reveals NF-kB activation in CLL mouse model LYMPHOID NEOPLASIA A Sleeping Beauty screen reveals NF-kB activation in CLL mouse model,* **121**, 4355 (2013).

[113] *MATLAB version 8.1.0.604 (R2013a)*, The Mathworks, Inc., Natick, Massachusetts (2013).

[114] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, D. Tax, and S. Verzakov, *PRTools 4.1, A Matlab Toolbox for Pattern Recognition*, Delft University of Technology. (2013).

[115] D. W. Huang, B. T. Sherman, and R. A. Lempicki, *Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists,* Nucleic Acids Research **37**, 1 (2009).

[116] D. W. Huang, R. a. Lempicki, and B. T. Sherman, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols **4**, 44 (2009).

[117] Y. S. Jo, M. S. Kim, N. J. Yoo, and S. H. Lee, *Frameshift Mutations of AKAP9 Gene in Gastric and Colorectal Cancers with HighMicrosatellite Instability,* Pathology and Oncology Research **21**, 181 (2016).

[118] K. Shinmura, T. Kahyo, H. Kato, H. Igarashi, S. Matsuura, S. Nakamura, K. Kurachi, T. Nakamura, H. Ogawa, K. Funai, M. Tanahashi, H. Niwa, and H. Sugimura, *RSPO fusion transcripts in colorectal cancer in Japanese population,* Molecular Biology Reports **41**, 5375 (2014).

[119] F. Mazurier, a. Fontanellas, S. Salesse, L. Taine, S. Landriau, F. Moreau-Gaudry, J. Reiffers, B. Peault, J. P. Di Santo, and H. de Verneuil, *A novel immunodeficient mouse*

model–RAG2 x common cytokine receptor gamma chain double mutants–requiring exogenous cytokine administration for human hematopoietic stem cell engraftment. Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research **19**, 533 (1999).

[120] H. Qian, N. Buza-Vidas, C. D. Hyland, C. T. Jensen, J. Antonchuk, R. Månsson, L. A. Thoren, M. Ekblom, W. S. Alexander, and S. E. W. Jacobsen, *Critical role of thrombopoietin in maintaining adult quiescent hematopoietic stem cells,* Cell Stem Cell **1**, 671 (2007).

[121] H. Yoshihara, F. Arai, K. Hosokawa, T. Hagiwara, K. Takubo, Y. Nakamura, Y. Gomei, H. Iwasaki, S. Matsuoka, K. Miyamoto, H. Miyazaki, T. Takahashi, and T. Suda, *Thrombopoietin/MPL signaling regulates hematopoietic stem cell quiescence and interaction with the osteoblastic niche,* Cell Stem Cell **1**, 685 (2007).

[122] J. Liu, S. Liu, M. Xia, S. Xu, C. Wang, Y. Bao, M. Jiang, Y. Wu, and T. Xu, *Rhomboid domain-containing protein 3 is a negative regulator of TLR3-triggered natural killer cell activation,* Proceedings of the National Academy of Sciences **110**, 7814 (2013).

[123] J. D. Riordan, V. W. Keng, B. R. Tschida, T. E. Scheetz, J. B. Bell, K. M. Podetz-Pedersen, C. D. Moser, N. G. Copeland, N. A. Jenkins, L. R. Roberts, D. A. Largaespada, and A. J. Dupuy, *Identification of Rtl1, a retrotransposon-derived imprinted gene, as a novel driver of hepatocarcinogenesis,* PLoS Genetics **9**, e1003441 (2013).

[124] X. Zhang, J. Liu, S. Yan, K. Huang, Y. Bai, and S. Zheng, *High expression of N-acetyltransferase 10 : a novel independent prognostic marker of worse outcome in patients with hepatocellular carcinoma,* International Journal of Clinical and Experimental Pathology **8**, 14765 (2015).

[125] S. Sarfraz, S. Hamid, S. Ali, W. Jafri, and A. a. Siddiqui, *Modulations of cell cycle checkpoints during HCV associated disease,* BMC infectious diseases **9**, 1 (2009).

[126] D. Wallis, M. Hamblen, Y. Zhou, K. J. T. Venken, A. Schumacher, H. L. Grimes, H. Y. Zoghbi, S. H. Orkin, and H. J. Bellen, *The zinc finger transcription factor Gfi1, implicated in lymphomagenesis, is required for inner ear hair cell differentiation and survival.* Development (Cambridge, England) **130**, 221 (2003).

[127] J. Erikson, a. Ar-Rushdi, H. L. Drwinga, P. C. Nowell, and C. M. Croce, *Transcriptional activation of the translocated c-myc oncogene in burkitt lymphoma.* Proceedings of the National Academy of Sciences of the United States of America **80**, 820 (1983).

[128] G. Meierhoff, U. Dehmel, H. J. Gruss, O. Rosnet, D. Birnbaum, H. Quentmeier, W. Dirks, and H. G. Drexler, *Expression of FLT3 receptor and FLT3-ligand in human leukemia-lymphoma cell lines.* Leukemia **9**, 1368 (1995).

[129] H. Ikeda, Y. Kanakura, T. Tamaki, a. Kuriu, H. Kitayama, J. Ishikawa, Y. Kanayama, T. Yonezawa, S. Tarui, and J. D. Griffin, *Expression and functional role of the proto-oncogene c-kit in acute myeloblastic leukemia cells.* Blood **78**, 2962 (1991).

[130] U. E. Höpken, H. D. Foss, D. Meyer, M. Hinz, K. Leder, H. Stein, and M. Lipp, *Up-regulation of the chemokine receptor CCR7 in classical but not in lymphocyte-predominant Hodgkin disease correlates with distinct dissemination of neoplastic cells in lymphoid organs,* Blood **99**, 1109 (2002).

[131] T. Kono and G. E. N. Yamada, *Murine interleukin 2 receptor j8 chain: Dysregulated gene expression in lymphoma line EL-4 caused by a promoter insertion,* Proceedings of the National Academy of Sciences **5**, 1806 (1990).

[132] C. J. de Boer, J. C. Kluin-Nelemans, E. Dreef, M. G. Kester, P. M. Kluin, E. Schuuring, and J. H. van Krieken, *Involvement of the CCND1 gene in hairy cell leukemia.* Annals of oncology : official journal of the European Society for Medical Oncology / ESMO **7**, 251 (1996).

[133] A. Pillai, M. Campbell, R. Levy, and A. M. Krensky, *Interleukin 3 is a growth factor for human follicular B cell lymphoma,* The Journal of experimental medicine **175**, 371 (1992).

[134] B. Raton, *Phenome-genome association studies of pancreatic cancer: new targets for therapy and diagnosis,* Cancer Genomics-Proteomics **12**, 9 (2015).

[135] E. J. Childs, E. Mocci, D. Campa, P. M. Bracci, S. Gallinger, M. Goggins, D. Li, R. E. Neale, S. H. Olson, G. Scelo, L. T. Amundadottir, W. R. Bamlet, M. F. Bijlsma, A. Blackford, M. Borges, P. Brennan, H. Brenner, H. B. Bueno-de Mesquita, F. Canzian, G. Capurso, G. M. Cavestro, K. G. Chaffee, S. J. Chanock, S. P. Cleary, M. Cotterchio, L. Foretova, C. Fuchs, N. Funel, M. Gazouli, M. Hassan, J. M. Herman, I. Holcatova, E. A. Holly, R. N. Hoover, R. J. Hung, V. Janout, T. J. Key, J. Kupcinskas, R. C. Kurtz, S. Landi, L. Lu, E. Malecka-Panas, A. Mambrini, B. Mohelnikova-Duchonova, J. P. Neoptolemos, A. L. Oberg, I. Orlow, C. Pasquali, R. Pezzilli, C. Rizzato, A. Saldia, A. Scarpa, R. Z. Stolzenberg-Solomon, O. Strobel, F. Tavano, Y. K. Vashist, P. Vodicka, B. M. Wolpin, H. Yu, G. M. Petersen, H. A. Risch, and A. P. Klein, *Common variation at 2p13.3, 3q29,*

7p13 and 17q25.1 associated with susceptibility to pancreatic cancer, Nature Genetics **47**, 911 (2015).

A

APPENDIX - DATA TRANSFORMATION

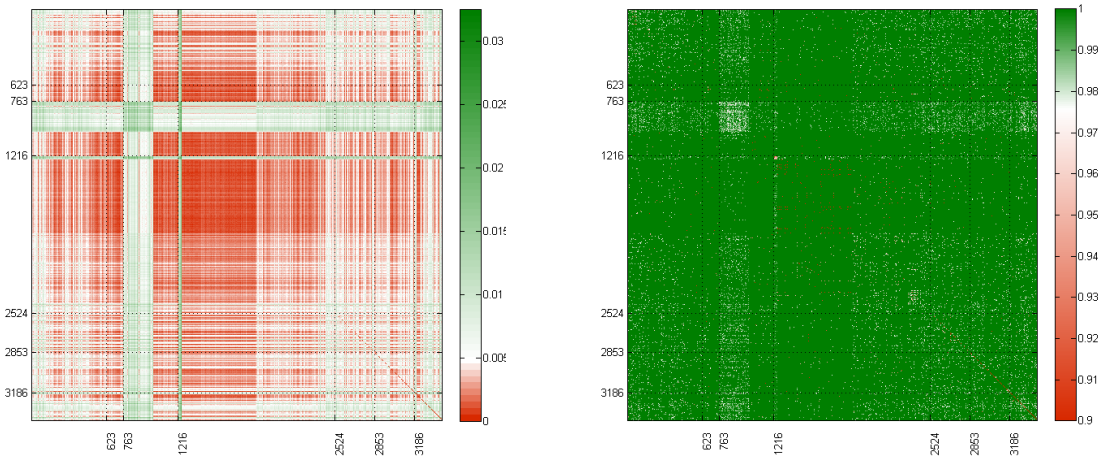
A.1. EXAMPLES OF DISTANCE METRICS

Hamming [69] and Jaccard distance [70] were selected to calculate distances between samples. Some examples of these distances are presented bellow.

$$
\begin{array}{ll}
S1 = \{1,1,1,1,1,1,1,1,1,1\} & Hamming: \ 0 \\
S2 = \{1,1,1,1,1,1,1,1,1,1\} & Jaccard: \ 0 \\
 & \\
S1 = \{0,0,0,0,0,0,0,0,0,0\} & Hamming: \ 0 \\
S2 = \{0,0,0,0,0,0,0,0,0,0\} & Jaccard: \ error \\
 & \\
S1 = \{1,1,1,1,1,0,0,0,0,0\} & Hamming: \ 1 \\
S2 = \{0,0,0,0,0,1,1,1,1,1\} & Jaccard: \ 1 \\
 & \\
S1 = \{1,1,1,1,1,1,1,1,1,1\} & Hamming: \ 0.5 \\
S2 = \{1,1,1,1,1,0,0,0,0,0\} & Jaccard: \ 0.5 \\
 & \\
S1 = \{0,0,0,0,0,0,0,0,0,0\} & Hamming: \ 0.5 \\
S2 = \{1,1,1,1,1,0,0,0,0,0\} & Jaccard: \ 1 \\
 & \\
S1 = \{1,1,1,0,0,0,0,0,0,0\} & Hamming: \ 0.4 \\
S2 = \{1,1,1,0,0,0,1,1,1,1\} & Jaccard: \ 0.57 \\
 & \\
S1 = \{1,1,1,1,0,0,0,0,0,0\} & Hamming: \ 0.2 \\
S2 = \{1,1,1,1,0,0,0,0,1,1\} & Jaccard: \ 0.33 \\
 & \\
S1 = \{1,1,0,0,0,0,0,0,0,0\} & Hamming: \ 0.2 \\
S2 = \{0,0,0,0,0,0,0,0,0,0\} & Jaccard: \ 1 \\
 & \\
S1 = \{0,0,1,1,1,1,1,1,1,1\} & Hamming: \ 0.2 \\
S2 = \{1,1,1,1,1,1,1,1,1,1\} & Jaccard: \ 0.2
\end{array}
\tag{A.1}
$$

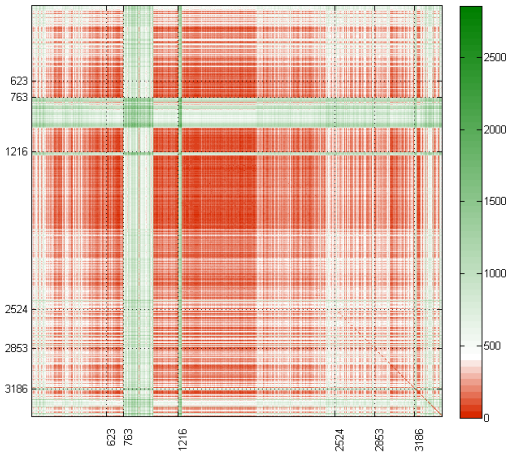
A.2. ENTIRE DATA TRANSFORMATION

The data wer transformed by the Hamming distance, Jaccard distance, gene dependent method (GDM) and gene independent method (GIM). This transformation creates a distance matrix between samples. Figure A.1 the heat map of those transformations.

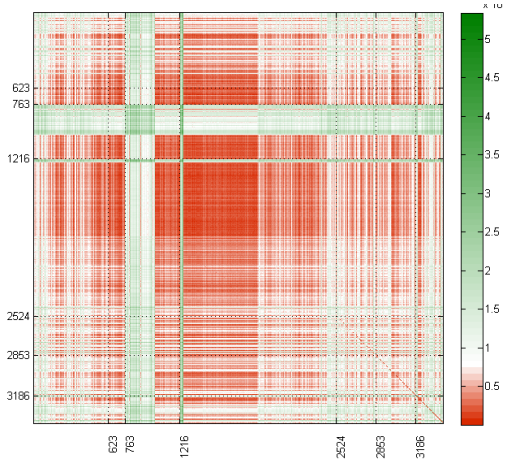


(a) Hamming distance

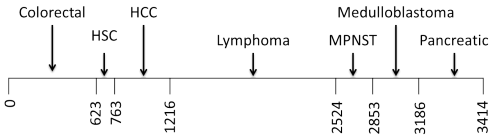
(b) Jaccard distance



(c) gene dependent method (GDM)



(d) gene independent method (GIM)



(e) Tumour types.

Figure A.1: Heat map of the distance matrices generated by the Hamming distance, Jaccard distance, gene dependent method (GDM) and gene independent method (GIM). Subfigure **e** shows the position of each tumour type.

B

APPENDIX - CROSS-VALIDATION VALUES

The values of Figure 6.3 are described in the following four tables:

Table B.1: Cross-validation using the Hamming transformation.
It was calculated the cross-validation, between two samples, using three classifiers: nearest mean classifier; *k*-nearest neighbour; and support vector machine.

Nearest Mean classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.1575	0.4236	0.2752	0.4555	0.4554	0.3492
HSC	0.1575	-	0.2227	0.3861	0.1479	0.1244	0.1001
HCC	0.4236	0.2227	-	0.2975	0.4279	0.4394	0.5028
Lymphoma	0.2752	0.3861	0.2975	-	0.2464	0.2380	0.1876
MPNST	0.4555	0.1479	0.4279	0.2464	-	0.4884	0.3837
Medulloblastoma	0.4554	0.1244	0.4394	0.2380	0.4884	-	0.3789
Pancreatic	0.3492	0.1001	0.5028	0.1876	0.3837	0.3789	-

k-Nearest Neighbour classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.1675	0.1612	0.2732	0.4559	0.4429	0.3721
HSC	0.1675	-	0.1907	0.3502	0.1501	0.1381	0.1076
HCC	0.1612	0.1907	-	0.1854	0.1642	0.1441	0.2492
Lymphoma	0.2732	0.3502	0.1854	-	0.2607	0.2398	0.2139
MPNST	0.4559	0.1501	0.1642	0.2607	-	0.4976	0.3861
Medulloblastoma	0.4429	0.1381	0.1441	0.2398	0.4976	-	0.3161
Pancreatic	0.3721	0.1076	0.2492	0.2139	0.3861	0.3161	-

Suport Vector Machine classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.1342	0.4205	0.2564	0.2602	0.3055	0.3307
HSC	0.1342	-	0.1761	0.2557	0.1479	0.1246	0.1001
HCC	0.4205	0.1761	-	0.2733	0.1770	0.1481	0.1942
Lymphoma	0.2564	0.2557	0.2733	-	0.1304	0.1376	0.1297
MPNST	0.2602	0.1479	0.1770	0.1304	-	0.4832	0.3812
Medulloblastoma	0.3055	0.1246	0.1481	0.1376	0.4832	-	0.3741
Pancreatic	0.3307	0.1001	0.1942	0.1297	0.3812	0.3741	-

Table B.2: Cross-validation using the Jaccard transformation.
It was calculated the cross-validation, between two samples, using three classifiers: nearest mean classifier; *k*-nearest neighbour; and support vector machine.

Nearest Mean classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.1683	0.1453	0.1300	0.1808	0.2456	0.2353
HSC	0.1683	-	0.0925	0.2368	0.2131	0.1748	0.0918
HCC	0.1453	0.0925	-	0.0681	0.1560	0.1335	0.1975
Lymphoma	0.1300	0.2368	0.0681	-	0.0895	0.1148	0.0698
MPNST	0.1808	0.2131	0.1560	0.0895	-	0.2767	0.1639
Medulloblastoma	0.2456	0.1748	0.1335	0.1148	0.2767	-	0.1682
Pancreatic	0.2353	0.0918	0.1975	0.0698	0.1639	0.1682	-

k-Nearest Neighbour classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.1958	0.0939	0.1384	0.2934	0.2630	0.3494
HSC	0.1958	-	0.2163	0.2556	0.1319	0.1149	0.1007
HCC	0.0939	0.2163	-	0.0913	0.1240	0.0748	0.1708
Lymphoma	0.1384	0.2556	0.0913	-	0.1168	0.1164	0.1162
MPNST	0.2934	0.1319	0.1240	0.1168	-	0.3057	0.2476
Medulloblastoma	0.2630	0.1149	0.0748	0.1164	0.3057	-	0.2630
Pancreatic	0.3494	0.1007	0.1708	0.1162	0.2476	0.2630	-

Suport Vector Machine classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.1886	0.0972	0.1296	0.1892	0.2484	0.2461
HSC	0.1886	-	0.0713	0.2606	0.1779	0.1608	0.0767
HCC	0.0972	0.0713	-	0.0506	0.0955	0.0895	0.1312
Lymphoma	0.1296	0.2606	0.0506	-	0.0880	0.1075	0.0683
MPNST	0.1892	0.1779	0.0955	0.0880	-	0.2706	0.1675
Medulloblastoma	0.2484	0.1608	0.0895	0.1075	0.2706	-	0.1752
Pancreatic	0.2461	0.0767	0.1312	0.0683	0.1675	0.1752	-

Table B.3: Cross-validation using the GDM transformation.
It was calculated the cross-validation, between two samples, using three classifiers: nearest mean classifier; *k*-nearest neighbour; and support vector machine.

Nearest Mean classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.1493	0.3940	0.3670	0.4595	0.5119	0.3618
HSC	0.1493	-	0.1955	0.2731	0.1400	0.1326	0.1069
HCC	0.3940	0.1955	-	0.3314	0.4005	0.3956	0.4547
Lymphoma	0.3670	0.2731	0.3314	-	0.3482	0.3686	0.2759
MPNST	0.4595	0.1400	0.4005	0.3482	-	0.4588	0.4030
Medulloblastoma	0.5119	0.1326	0.3956	0.3686	0.4588	-	0.3569
Pancreatic	0.3618	0.1069	0.4547	0.2759	0.4030	0.3569	-

k-Nearest Neighbour classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.1828	0.1771	0.3574	0.4473	0.4636	0.3749
HSC	0.1828	-	0.1915	0.3161	0.1534	0.1496	0.1169
HCC	0.1771	0.1915	-	0.1519	0.1641	0.1338	0.2182
Lymphoma	0.3574	0.3161	0.1519	-	0.2576	0.3306	0.2254
MPNST	0.4473	0.1534	0.1641	0.2576	-	0.4900	0.4273
Medulloblastoma	0.4636	0.1496	0.1338	0.3306	0.4900	-	0.3718
Pancreatic	0.3749	0.1169	0.2182	0.2254	0.4273	0.3718	-

Suport Vector Machine classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.3700	0.3864	0.4011	0.5086	0.5373	0.4463
HSC	0.3700	-	0.4098	0.4310	0.4127	0.3922	0.4261
HCC	0.3864	0.4098	-	0.3188	0.4400	0.3448	0.4441
Lymphoma	0.4011	0.4310	0.3188	-	0.3998	0.3986	0.4026
MPNST	0.5086	0.4127	0.4400	0.3998	-	0.4682	0.5356
Medulloblastoma	0.5373	0.3922	0.3448	0.3986	0.4682	-	0.4196
Pancreatic	0.4463	0.4261	0.4441	0.4026	0.5356	0.4196	-

Table B.4: Cross-validation using the GIM transformation.
It was calculated the cross-validation, between two samples, using three classifiers: nearest mean classifier; *k*-nearest neighbour; and support vector machine.

Nearest Mean classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.1526	0.4230	0.2748	0.4487	0.4545	0.3510
HSC	0.1526	-	0.2234	0.3843	0.1462	0.1251	0.0935
HCC	0.4230	0.2234	-	0.2974	0.4299	0.4418	0.5033
Lymphoma	0.2748	0.3843	0.2974	-	0.2483	0.2387	0.1900
MPNST	0.4487	0.1462	0.4299	0.2483	-	0.4903	0.3910
Medulloblastoma	0.4545	0.1251	0.4418	0.2387	0.4903	-	0.3753
Pancreatic	0.3510	0.0935	0.5033	0.1900	0.3910	0.3753	-

k-Nearest Neighbour classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.1607	0.1598	0.2697	0.4510	0.4492	0.3538
HSC	0.1607	-	0.1836	0.3513	0.1476	0.1389	0.1258
HCC	0.1598	0.1836	-	0.2032	0.1583	0.1461	0.2530
Lymphoma	0.2697	0.3513	0.2032	-	0.2592	0.2472	0.2059
MPNST	0.4510	0.1476	0.1583	0.2592	-	0.4705	0.3728
Medulloblastoma	0.4492	0.1389	0.1461	0.2472	0.4705	-	0.3458
Pancreatic	0.3538	0.1258	0.2530	0.2059	0.3728	0.3458	-

Suport Vector Machine classification

	Colorectal	**HSC**	**HCC**	**Lymphoma**	**MPNST**	**Medulloblastoma**	**Pancreatic**
Colorectal	-	0.4433	0.4106	0.4941	0.5008	0.5005	0.4932
HSC	0.4433	-	0.4174	0.4475	0.4348	0.5973	0.3949
HCC	0.4106	0.4174	-	0.4383	0.3748	0.6201	0.3651
Lymphoma	0.4941	0.4475	0.4383	-	0.4858	0.4750	0.5040
MPNST	0.5008	0.4348	0.3748	0.4858	-	0.4782	0.4776
Medulloblastoma	0.5005	0.5973	0.6201	0.4750	0.4782	-	0.4997
Pancreatic	0.4932	0.3949	0.3651	0.5040	0.4776	0.4997	-

C

APPENDIX - GENE LIST

The list of predicted genes involved in a specific tumour type are described bellow:

1110059e24Rik RIKEN cDNA 1110059E24 gene; predicted gene 9504

1810049j17Rik RIKEN cDNA 1810049J17 gene

2310061n02Rik RIKEN cDNA 2310061N02 gene

3110070m22Rik RIKEN cDNA 3110070M22 gene

4931406c07Rik RIKEN cDNA 4931406C07 gene

4931422a03Rik RIKEN cDNA 4931422A03 gene

4933406p04Rik RIKEN cDNA 4933406P04 gene

4933440n22Rik RIKEN cDNA 4933440N22 gene

Abhd13 abhydrolase domain containing 13

Akap9 A-kinase anchor protein 9

Akr1c20 aldo-keto reductase family 1, member C20

Alg9 asparagine-linked glycosylation 9 homolog (yeast, alpha 1,2 mannosyltransferase)

Arid4a AT rich interactive domain 4A (RBP1-like)

Armc7 armadillo repeat containing 7

Ascl2 achaete-scute family bHLH transcription factor 2

Atp5h ATP synthase, H+ transporting, mitochondrial Fo complex, subunit d

Atxn7 ataxin 7

Baz2a bromodomain adjacent to zinc finger domain, 2A

Bc003331 cDNA sequence BC003331

Bcl9 B-cell CLL/lymphoma 9

Cbx3 chromobox homolog 3

Cbx5 ataxin 7

Ccdc138 coiled-coil domain containing 138

Ccl25 C-C motif chemokine 25

Ccnd1 cyclin D1

Ccr7 chemokine (C-C motif) receptor 7

Cdkl3 cyclin-dependent kinase-like 3

Cmya5 cardiomyopathy associated 5

Csf2 colony stimulating factor 2

Cyp2j8 cytochrome P450, family 2, subfamily j, polypeptide 8

Ddx25 DEAD (Asp-Glu-Ala-Asp) box polypeptide 25

Dnajb4 DnaJ (Hsp40) homolog, subfamily B, member 4

Epor erythropoietin receptor

Eras ES cell-expressed Ras

Fam179b Family with sequence similarity 179, member B

Fdxacb1 ferredoxin-fold anticodon binding domain containing 1

Flt3 fms-related tyrosine kinase 3

Foxr2 forkhead box R2

Fuca1 fucosidase, alpha-L- 1, tissue

Gfi1 growth factor independent 1 transcription repressor

Gm10337 predicted gene 10337

Gm10537 predicted gene 10537

Gm10542 predicted gene 10542

Gm10638 predicted gene 10638

Gm10974 predicted gene 10974

Gm11273 predicted gene 11273

Gm17535 predicted gene 17535

Gm26965 predicted gene 26965

Gm5129 predicted gene 5129

Gpr75 G protein-coupled receptor 75

Gtf2h2 general transcription factor II H, polypeptide 2

Hdlbp high density lipoprotein (HDL) binding protein

Hecw1 HECT, C2 and WW domain containing E3 ubiquitin protein ligase 1

Hus1b Hus1 homolog b (S. pombe)

Ift46 intraflagellar transport 46

Il2rb interleukin 2 receptor, beta

Il3 interleukin 3

Kit kit oncogene

Kmt2e lysine (K)-specific methyltransferase 2E

Kntc1 kinetochore associated 1

Lig4 ligase IV, DNA, ATP-dependent

Matr3 matrin 3; similar to Matrin 3

Mrpl48 mitochondrial ribosomal protein L48

Myc myelocytomatosis oncogene

Naip2 NLR family, apoptosis inhibitory protein 2

Nat10 N-acetyltransferase 10

Pate2 prostate and testis expressed 2

Pcdhga5 protocadherin gamma subfamily A, 5

Pcdhgb2 protocadherin gamma subfamily B, 2

Pdcd11 programmed cell death 11

Pik3r5 phosphoinositide-3-kinase, regulatory subunit 5

Plod1 procollagen-lysine, 2-oxoglutarate 5-dioxygenase 1

Prl3c1 prolactin family 3, subfamily c, member 1

Prr36 proline rich 36

Rag2 recombination activating gene 2

Rapgefl1 Rap guanine nucleotide exchange factor (GEF)-like 1

Rft1 RFT1 homolog (S. cerevisiae)

Rhbdd3 rhomboid domain containing 3

Rspo2 R-spondin 2 homolog (Xenopus laevis)

Rspry1 ring finger and SPRY domain containing 1

Rtl1 retrotransposon-like 1; RIKEN cDNA 6430411K18 gene

Ryk receptor-like tyrosine kinase

Samd4 sterile alpha motif domain containing 4

Sepp1 selenoprotein P, plasma, 1

Slc6a18 solute carrier family 6 (neurotransmitter transporter), member 18

Snx9 sorting nexin 9

Spink11 serine peptidase inhibitor, Kazal type 11

Sugct succinyl-CoA:glutarate-CoA transferase

Tbr1 T-box brain gene 1

Tcte2 t-complex-associated testis expressed 2

Tgif2 TGFB-induced factor homeobox 2

Thpo thrombopoietin

Tmc1 transmembrane channel-like gene family 1

Tmem106a transmembrane protein 106A

Trove2 TROVE domain family, member 2

Trpm8 transient receptor potential cation channel, subfamily M, member 8

Tshb thyroid stimulating hormone, beta subunit

Ttc4 tetratricopeptide repeat domain 41

Vmn1r200 vomeronasal 1 receptor, H3

Xpnpep3 X-prolyl aminopeptidase (aminopeptidase P) 3, putative

Zfp106 zinc finger protein 106

Zfp53 zinc finger protein 53

Zfp87 zinc finger protein 87